Prediction of Daily Streamflow Using Various Kernel Function Based Regression: A Case Study in India

Manish Kumar

G B Pant University of Agriculture and Technology: Govind Ballabh Pant University of Agriculture & Technology

Ahmed Elbeltagi (✉ ahmedelbeltagy81@mans.edu.eg)

Mansoura University Faculty of Agriculture

Ankur Srivastava

Newcastle University

Anuradha Kumari

G B Pant University of Agriculture and Technology: Govind Ballabh Pant University of Agriculture & Technology

Rawshan Ali

–

Chaitanya Pande

AICRP

Tarate Suryakant Bajirao

//////

Abu Reza Md. Towfiqul Islam

National Institute of Disaster Management

Daniel Prakash Kushwaha

G B Pant University of Agriculture and Technology: Govind Ballabh Pant University of Agriculture & Technology

Research Article

Keywords: Stage-Discharge Modeling, Gaussian Functions, PUK Kernel, Govindpur

Posted Date: August 13th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-784271/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Prediction of Daily Streamflow using various kernel function based regression: A case study in India

Manish Kumar^{1*}, Ahmed Elbetagi ^{2,*}, Ankur Srivastava³, Anuradha Kumari ¹, Rawshan Ali⁴, Chaitanya Pande⁵, Tarate Suryakant Bajirao⁶, Abu Reza Md. Towfiqul Islam ⁷, Daniel Prakash Kushwaha¹

¹ Department of Soil and Water Conservation Engineering, College of Technology, G.B.P.U.A.&T., Pantnagar, Uttarakhand-263145, India; anuradhakhushio7@gmail.com; danielprakash1991@gmail.com

² Agricultural Engineering Dept., Faculty of Agriculture, Mansoura University, Mansoura 35516, Egypt; ahmedelbeltagy81@mans.edu.eg

³ School of Engineering, The University of Newcastle, Callaghan, NSW 2308, Australia

⁴ Departments of Petroleum, Koya Technical Institute, Erbil Polytechnic University, Erbil 44001, Kurdistan, Iraq

⁵ AICRP for Dryland Agriculture, Dr. Panjabrao Deshmukh Krishi Vidyapeeth, Akola, Maharashtra, India, chaitanay45@gmail.com

⁶ Department of Soil Science and Agriculture Chemistry, School of Agriculture, Lovely Professional University, Phagwara 144411, India; suryakant.26256@lpu.co.in

⁷ Department of Disaster Management, Begum Rokeya University, Rangpur 5400, Bangladeshi, towfiq_dm@brur.ac.bd

* Correspondence: Manish Kumar (manishgbpuat@yahoo.com); Ahmed Elbetagi (ahmedelbeltagy81@mans.edu.eg).

Abstract

River daily discharge estimation and modeling considers an important step for scheduling and planning different water resources for sustainable socio-economic development. In the current work, four techniques of Gaussian processes regression (GPR): Polynomial Kernel, Radial Basis Function Kernel, Normalized Polynomial Kernel, and PUK Kernel, were used to model the daily discharge. Hydrological-datasets containing daily-stage (m) and discharge (m^3/sec) were gathered over the period from 2004-2013. The datasets were divided into two sections: (i) models training containing 70% (2004-2010) of the total data and (ii) remaining 30% (2011-2013) were for testing. Comparing all the four developed models, our findings show that the superlative model was the PUK-Kernel model with a correlation coefficient (r) of 0.96, MAE of $36.70 \text{ m}^3/\text{s}$, RMSE of $90.92 \text{ m}^3/\text{s}$, RAE of 17.50 %, RRSE of 26.05 % in the training period. Whereas, it performed equally well in the testing period with $r = 0.97$, MAE = $44.84 \text{ m}^3/\text{s}$, RMSE = $95.05 \text{ m}^3/\text{s}$, RAE = 17.98 %, RRSE = 24.94 % in the testing period. Our findings can be included that GPR-PUK was more accurate and stable than other models, and can be used to help water-users, decision-makers, development-planners for managing water resources and achieving sustainable development.

Keyword: Stage-Discharge Modeling, Gaussian Functions, PUK Kernel, Govindpur

1. Introduction

River discharge is a critical component of the water cycle since it affects water supply, irrigation, and flood prediction. River discharge and water level observation is an important research point in hydrological and hydraulic modeling, and it also represents a piece of vital knowledge for managing and monitoring water sources. For example, the precise stage-discharge calculation is critical for estimating design flows for different hydraulic infrastructures such as bridges, culverts, and canals (Manish et al, 2020). Direct flow discharge measurements are most often complicated or not possible in very complex or compound rivers. In addition, neither discharge nor water level may be accessible in certain cases or have the same data series record. Flow rating curves (FRCs) are therefore the normal and most general method for estimating missing details about a particular variable in such circumstances (Srivastava et al., 2018). With more than just a handful, FRCs has been based on calibrated historical phase-discharge rating records for decades (i.e., discharge and water level) (Mohantyet al. 2019). FRCs is most commonly used for medium and large rivers, where time and money can be expensive to make direct measurements. In contrast to FRCs, both flow discharge and water level can be directly determined in small rivers using current meters or other specialized technologies. The performance of the FRCs, on the other hand, can be influenced by river stage geometry and measurement uncertainty in general, limiting the estimation of high values (Manfreda, S.et al, 2020).

For non-linear regression and classification problems, Gaussian process models have become extremely common. They are expressive, interpretable, resist overfitting, and have excellent predictive success in a variety of rigorous empirical comparisons (Rasmussen, 1996; Kuss and Rasmussen, 2005; Rasmussen and Williams, 2006).

Gaussian process regression was established from research into neural networks in machine learning. Neal (1996) demonstrated that Bayesian neural networks were Gaussian processes when the number of hidden units reached infinitely and hypothesized that there could be easier ways to allow inference in this situation. The foundation of subsequent Gaussian process models became these basic inference techniques. Neural networks, however, were partly inspired by their ability to capture associations between multiple outputs (responses), using hidden adaptive units that were shared between the outputs. This skill was lost at the infinity limit. There has recently been a surge of interest in extending the Gaussian process regression model to account for associations between output variables (Alvarez and Lawrence, 2011; Yu et al., 2009; Alvarez and Lawrence, 2008; Bonilla et al., 2008; Osborne et al., 2008; Teh et al., 2005; Boyle and Frean, 2004). These are also referred to as models of 'multi-task learning or 'multiple outputs' regression. To make better predictions, it can be used to capture correlations between outputs (response variables). Suppose we want to predict the concentration of cadmium in the Swiss Jura region, where geologists are interested in the concentration of heavy metals. A regular regression model of the Gaussian process will only be capable of using measurements from cadmium preparation. Related heavy metal calculations can be combined with a multi-task approach to improve cadmium predictions (Goovaerts, 1997). If we make use of how these (signal) associations change with geographical position, we may further improve predictions.

Similarly, extending Gaussian process (GP) regression to allow for input-dependent noise variances has piqued researchers' interest (Goldberg et al., 1998; Kersting et al., 2007; Adams and Stegle, 2008; Turner and Sahani, 2008; Turner, 2010; Wilson and Ghahramani, 2010a,b; Lazaro-Gredilla and Titsias, 2011). Wilson and Ghahramani (2010b, 2011) and Fox and Dunson (2011) extended the GP framework to handle input dependent noise associations between

multiple-output (response) variables; GP regression with non-stationary covariance function amplitudes (Turner and Sahani, 2008; Adams and Stegle, 2008) and length-scales (Gibbs, 1997; Schmidt and O'Hagan, 2003) and strong tailed predictive distributions for outlier rejection (Neal, 1997; Vanhatalo et al., 2009) are other extensions (De Finetti, 1956; Dawid, 1973; O'Hagan, 1979). We provide a unifying view of sparse approximations for GP regression in this paper. Our approach is simple but powerful: we evaluate the posterior for each algorithm and calculate the effectiveness before which it is used. As a result, we re-interpret the algorithms as "exact inference with an approximate prior," rather than the existing (common) interpretation of "approximate inference with the exact prior". In terms of prior assumptions about the function, this method has the advantage of expressing the approximations explicitly, which makes the implications of the approximations much simpler to grasp. Although our view of approximations is not the only one possible, it has the advantage of bringing under one umbrella all current probabilistic sparse approximations, thereby allowing a direct comparison and disclosing the relationship between them. We introduce Gaussian Process Regression (GPR) in this article, a modern regression approach that incorporates the structural properties of Bayesian neural networks (Neal, 1996) with Gaussian processes. This network is an adaptive combination of Gaussian systems that accommodates input-based signal and noise associations between several output variables, input dependent length scales and amplitudes, and strong tailed predictive distributions without requiring costly or numerically inefficient computations. Four Gaussian processes regression (GPR): Polynomial Kernel, Radial Basis Function Kernel, Normalized Polynomial Kernel, and PUK Kernel, were applied for daily discharge modeling. Therefore, the objectives of this study are to 1) fit four GPR models for modeling daily streamflow under different five scenarios, 2) compare the accuracy and stability of these developed models in

streamflow prediction, and 3) select the best outcomes have been provided from the created GPR models based on the performance metrics.

2. Materials and Methods

2.1. Study area

The research field includes the NH-5 road bridge Govindpur, also known as Govindpur, is a town in Balasore district of Orissa, India, with latitude $21^{\circ}32'52''$ N and longitude $86^{\circ}55'14''$ E. The study site is situated on the Burhabalang River, which is an east-flowing river that is also a tributary of the Subarnarekha river basin in Orissa State (Fig. 1). The drainage basin's contributing area is 4495 km^2 . The India-Water Resources Information System (WRIS) portal provided hydrological data for 10 years (1 June 2004–31 October 2013), including daily stage (m) and discharge (m^3/sec). Fig. 2 shows a time series plot of the overall available datasets of stage and discharge versus time. These datasets were divided into two parts: (i) training dataset containing 70% of the total data (1 June 2004 to 31 October 2010) that was used for models training, and (ii) testing dataset containing 30% of the total data (1 June 2011 to 31 October 2013) that was used to validate the prediction capability of the implemented models. The statistical analysis of given dataset of stage (m) and discharge (cumec) are presented in Table 1. The statistical analysis includes, mean, standard deviation (Std. Dev.), minimum, quartile 1 (Q1), median, quartile 3 (Q3) and maximum values. It gives basic statistical information about training data and testing data to get an idea about the nature of different datasets.

Fig. 1. Location map of study area

Fig. 2. Time series plot of stage and discharge dataset of study area

Table 1 Statistical analysis for Stage (m) and Discharge (cumec) parameters of study area

Variable	Mean	Std. Dev	Minimum	Q1	Median	Q3	Maximum
Training							
Stage	2.946	1.581	0.860	1.840	2.520	3.500	8.840
Discharge	243.5	349.2	1.4	61.2	136.8	265.0	2885.9
Testing							
Stage	2.754	1.703	0.8600	1.480	2.200	3.440	9.240
Discharge	291.6	381.5	3.6	77.8	157.1	342.4	2685.6

2.2. Machine learning methods used: Gaussian processes regression

Gaussian processes are modern supervised machine learning models utilized in regression and classified process (Williams and Rasmussen, 1996; Rasmussen et al. 2012). It is important in the field of machine learning models. GPR mainly knows that the learning sample tracks the previous chances of a Gaussian regression. It calculates the consistent probability afterwards. It is

produced according to the Bayesian model of linear regression. To describe the covariance of a previous scattering over the target functions, GPR uses the kernel to explain a chance function with the detected training information. A (Gaussian) post circulation over target functions, whose average is used for the data prediction, is explained according to the theorem of Bayes.

The gathering of converted forecasters and predictands form a column vector $z^T = [z_1, z_2, \dots, z_d]$. $z_i \sim N(\mu_i, \sigma_i)$ d is the dimension of forecaster and predictands, the combined Gaussian distribution for the d variables

$$z \sim N(\mu, \Sigma)$$

$$\mu^T = [\mu_1, \mu_2, \dots, \mu_d] \quad (1)$$

$$\Sigma = \text{diag}(\sigma) X R X \text{diag}(\sigma) \quad (2)$$

Where $\text{diag}(\sigma)$ is a diagonal matrix from the standard deviation, $\sigma^T = [\sigma_1, \sigma_2, \dots, \sigma_d]$, and R is the proportional correlation matrix.

Random processes represent Gaussian processes, where the joint distribution of a random variable is typical for the interconnection of many transformed forecasts and predictors following Gaussian distribution. Mean vector, standard vector and correspondence matrix; (Sun et al. 2014).

Rasmussen and its staff Gaussian technique of regression initially introduced 'the principled, virtually and mathematically based on the approach to kernel creating' (GPR) (Rasmussen et al. 2006). GPR's advantage over numerous other advantages lies in seamlessly integrating hyperparameter approximation, model training and insecurity evaluations. Gaussian (GP) methods presume that Gaussian mixed distribution of the possibility of model output (Markonis et al. 2015).

Polynomial Kernel

In machine learning language, the polynomial kernel (poly kernel) is a kernel function that is commonly utilized with the support-vectors machines (SVMs) and other kernel-coded models which refer to parallel vector. The polynomial kernel automatically appears in order to achieve its similarity not only on the specified functions of the input samples, but also in their mixtures. Such groupings were classified as interaction characteristics in the sense of regression analysis. In the sense of regression analysis, these groups have been classified as characteristics of interaction. The contained polynomial kernel space is equal to a polynomial regression, but it is an educated sum of parameters without a combined blow-up. If the function data input is binary, then the functionality connects to logical function connections (Yoav Goldberg et al. 2008).

For degree-d polynomials, the polynomial kernel is well-defined as

$$K(x, Y) = (x^T, y + C)^d \quad (3)$$

where x and y are vectors in the input space, i.e., vectors of features calculated from exercise or trial samples and $c \geq 0$ is an unrestricted parameter trading off the approach of higher-order versus lower-order definitions in the polynomial. When $c = 0$, the kernel is known as homogeneous.

PUK Kernel

The Pearson Universal Kernel is an effective use of machine learning programming and helps to broadly interpret and understand the various fields of data. In particular this kernel formula is a highly accurate climate- and hydrological prediction. The general form of the Pearson VII function for curve appropriate obtained is assumed by

$$f(x) = \frac{H}{\left[1 + ((2(x-x_0)\sqrt{2^{(\frac{1}{\omega}-1)/\delta})^2})\right]} \omega \quad (4)$$

where H is the top tallness at the center x_0 of the peak, and x denotes the self-determining variable. The parameters σ and ω regulator the half-width and the following factor of the peak. However, a function ω goes to the class of effective kernel functions. Individual if its symbolic and positive semi-definite kernel matrix is consistent. To show that these situations are certainly being resolved by PUK, Uestuen redrafted Eq. (4) into a function of both vectors (Uestuen et al., 2006):

$$K(x_i, x_j) = \frac{1}{\left[1 + ((2\sqrt{|x_i - x_j|^2} \sqrt{2^{(1/\omega)-1}/\sigma})\right]}^{2\omega} \quad (5)$$

The PUFK has the same or more powerful plotting effect as with the standard functions of the kernel, which mainly contribute to the equal or better generalization of SVMs. The Weka package was used for the performance of algorithms (Inamdar et al., 2004) and the following parameters were defined as defaults in total.

Normalized Polynomial Kernel

In machine learning and data mining, standardized polynomial kernel graders are used. It can be used in different fields to predict data. A machine learning technology is now being used in large-scale analysis with AI, drones and robotic technology built on the programming of machine learning. The metrological and the remote sensing data is a huge data and no precise information by conventional methods, but more accurate information can be given in comparison with machine learning with kernel classification. (Scholkopf 1997).

The normalization polynomial kernel of the vectors of the initial space can be agreed as the most basic kind of preprocessing $x \in R^N$ is an input vector, the consistent normalized vector $\underline{x}$ can be described as :

$$\underline{x} = \frac{x}{\|x\|} \in R^N \quad (6)$$

Where $\|x\|^2 = \sum_{i=1}^N x_i^2$, this vector lies on a unit hypersphere of R^N . Feature space data normalization does not mean tough pre-processing because input vectors are not obviously useful, but can be seen as a kernel example of the above considered pre-processing, which is an extension of the normalization feature. Normalization in a function space basically involves trying to define the functions of the SVM kernel, because it works for unrefined input vectors (Scholkopf et al. 1999; Cortes et al. 1995). Accepted the $K(x, y)$ is the kernel function on behalf of a dot product in the feature space (W. Maass 2000). Normalization in the feature space then number to describing a new kernel function $\underline{K}(x, y)$ as respectively:

$$\underline{K}(x, y) = \frac{K(x, y)}{\sqrt{K(x, y)K(x, y)}} \in R \quad (7)$$

RBF kernel

The RBF kernel is a common kernel function that is used in various machine training kernel learning algorithms. It is widely used in vector machinery classification. Usually the RBF kernel is a good first option. This kernel maps samples to higher dimensional areas, which means that a nonlinear relationship of class labels and attributes can be dealt with as opposed to the linear kernel (Keerthi et al. 2013). In addition, RBF is a different situation in the linear kernel Normalized poly kernel, since the linear kernel with a penalty parameter $\underline{C}$ has performs with parameters the same as the RBF kernel. (C, y) Furthermore, in certain parameters the sigmoid kernel is like RBF One important idea is $0 < K_{ij} \leq 1$

In comparison, kernel values can be used to $(\gamma X_i^T + r > 1)$ for polynomial kernelsto infinity as long as the grade is high (Lin et al. 2003). Furthermore, in such parameters the sigmoid kernel is not true, i.e. not an internal two-vector product. In certain situations, the RBF kernel is not suitable. In particular, when the number of functions is very high, the linear kernel can be used (Vapnik 1995).

2.3. Methodology Adopted

The hydrological datasets containing the daily stage (m) and discharge (m³/sec) of 10 years (1st June 2004-31st October 2013) were used for analysis process. The output variable is current day discharge. Five different scenarios were used by lagging of stage and discharge data. Table 2 shows the input combinations used in each scenario to developed the models. Table 2 shows five days lag of stage data $h_t, h_{t-1}, h_{t-2}, h_{t-3}, h_{t-4}, h_{t-5}$ as present day, previous one day, two day lag, three day lag, four day lag and five day lag, respectively while five days lag discharge data $Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4}, Q_{t-5}$ as previous one day, two day lag, three day lag, four day lag and five day lag, respectively. Q_t is the present day discharge which is an output of the model.

Different techniques of gaussian function were used to predict present day discharge at Govindpur site. The techniques of gaussian function were Polynomial Kernel, Radial Basis Function Kernel, Normalized Polynomial Kernel and PUK Kernel. The obtained results of this model were evaluated by correlation coefficient, MAE, RMSE, RAE, RRSE. The flowchart methodology for discharge estimation is given in Fig. 3.

268 **Table 2** Input combination used in each scenario of study

Scenarios	Lagging	Input Combinations
Scenario-1	One day lag	h_t, h_{t-1}, Q_{t-1}
Scenario-2	Two day lag	$h_t, h_{t-1}, h_{t-2}, Q_{t-1}, Q_{t-2}$
Scenario-3	Three day lag	$h_t, h_{t-1}, h_{t-2}, h_{t-3}, Q_{t-1}, Q_{t-2}, Q_{t-3}$
Scenario-4	Four day lag	$h_t, h_{t-1}, h_{t-2}, h_{t-3}, h_{t-4}, Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4}$
Scenario-5	Five day lag	$h_t, h_{t-1}, h_{t-2}, h_{t-3}, h_{t-4}, h_{t-5}, Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4}, Q_{t-5}$

269
270 **Fig. 3.** Flowchart methodology of discharge estimation at study area

2.4. Performance indicators

Actual data of discharge (Q) and predicted values were compared using different performance evaluators in this study. To evaluate the accuracy of models, the following performance indicators have been selected using coefficient of determination (r), mean absolute error (MAE), root mean square error (RMSE), relative absolute error (RAE) and root relative square error (RRSE) (Malone et al. 2017; Elbeltagi et al. 2020a, b, c; Srivastava et al., 2017, 2018, 2020a, 2020b; Kumar et al., 2020 Kumari et al., 2020). All parameters defined as follows: D_A^i is an observed or actual value, Q_P^i is simulated or foreseen value, Q^- is the mean value of reference samples, and N is the total number of data points. The performance of developed models was also compared using visual inspection such as line diagram and scatter plot.

Correlation coefficient (r)

Correlation coefficient value ranges from -1 to +1 and the value close to +1 represents best fit. It is to measure collinearity between the actual (A) and predicted (P) variables in hydrological studies. It is oversensitive to extreme value (Kumar et al., 2020; Tarate et al., 2021).

$$r = \left[\frac{\sum_{i=1}^n \{(Q_A - \bar{Q}_P)(Q_A - \bar{Q}_P)\}}{\sqrt{\sum_{i=1}^n (Q_A - \bar{Q}_P)^2} \sqrt{\sum_{i=1}^n (Q_A - \bar{Q}_P)^2}} \right] \quad (8)$$

Mean Absolute Error

The mean absolute error evaluates the mean magnitude of the errors in a set of forecasts, without considering their sign. It's an average of the absolute differences between expected and observed values over the test sample. It is defined as:

$$MAE = \frac{1}{N} \sum_{i=1}^N |Q_P^i - Q_A^i| \quad (9)$$

Root mean square error

RMSE is the sample standard deviation of the differences between foreseen and actual values. It is given by:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (Q_A^i - Q_P^i)^2} \quad (10)$$

Relative Absolute Error

The relative absolute error takes the total absolute error and normalizes it by dividing by the total absolute error of the simple predictor.

$$\text{RAE} = \left| \frac{Q_A^i - Q_P^i}{Q_P^i} \right| \times 100 \quad (11)$$

Root Relative Squared Error

The relative squared error takes the total squared error and normalizes it by dividing by the total squared error of the simple predictor. By taking the square root of the relative squared error one reduces the error to the same dimensions as the quantity being predicted.

$$\text{RRSE} = \frac{\sqrt{\sum_{i=1}^N (Q_P^i - Q_A^i)^2}}{\sqrt{\sum_{i=1}^N (Q_A^i - Q^-)^2}} \quad (12)$$

3. Results and Discussion

In this section, we discuss about the quantitative and qualitative evaluation of the results obtained from the four different models namely Poly Kernel, RBF-Kernel, and Normalized. Poly Kernel and PUK-Kernel were used in this study for five different scenarios (scenarios 1-5). Among these scenarios which one is the best suitable and deciphering the role, each scenario is described below.

3.1. Evaluation of different models using statistical indicators

Table 3 shows statistical indicators by using five different error measures that are r , MAE, RMSE, RAE, RRSE for all four models, and four training and testing periods for scenario 1. Among all the four models the best model which simulated the observed streamflow discharge performance well can be seen of the PUK-Kernel model with $r = 0.96$, MAE = 36.70 m³/s, RMSE = 90.92 m³/s, RAE = 17.50 %, RRSE = 26.05 % in the training period. Whereas, it performed equally well in the testing period with $r = 0.97$, MAE = 44.84 m³/s, RMSE = 95.05 m³/s, RAE = 17.98 %, RRSE = 24.94 % in the testing period. However, among the selected models the least performance of simulated discharge was shown by Norm. Poly Kernel having $r = 0.78$, MAE = 131.67 m³/s, RMSE = 218.55 m³/s, RAE = 62.78 %, RRSE = 62.62 % in the training period. Whereas, it in the testing period the efficiencies slightly decreased with $r = 0.48$, MAE = 206.53 m³/s, RMSE = 334.28 m³/s, RAE = 82.84 %, RRSE = 87.72 % in the testing period. The order of model performance and capturing of the simulated discharge against observed discharge based on r from very good to unsatisfactory was attained PUK-kernel (0.97)> Poly Kernel (0.92)> RBF-Kernel (0.91)> Norm. Poly Kernel (0.48).

Table 3. Statistical indicators and error measures between observed and simulated discharge of four different models utilized in this study for scenario 1.

Model	Dataset	r	MAE m ³ /s	RMSE m ³ /s	RAE (%)	RRSE (%)
Poly Kernel	Training	0.8916	157.149	201.542	74.928	57.749
	Testing	0.927	223.454	262.531	89.632	68.894
RBF-Kernel	Training	0.875	88.549	201.614	42.22	57.770
	Testing	0.9138	136.207	241.647	54.635	63.414

Norm. Poly Kernel	Training	0.783	131.678	218.550	62.783	62.624
	Testing	0.4836	206.534	334.28	82.845	87.72
PUK-Kernel	Training	0.967	36.703	90.927	17.500	26.054
	Testing	0.97	44.847	95.057	17.989	24.945

335

336

Fig. 4. Reproduction of observed and simulated discharge for four different models as a function of time
a) Poly Kernel, b) RBF-Kernel, c) Norm. Poly Kernel, and d) PUK-Kernel for scenario 1

3.2. Time series analysis of the different models for different scenarios

Fig. 4 shows the results of the four different models (Fig. 4 a-d) for scenario 1 plotted between observed and simulated discharge hydrograph in the form of time series and scatter plots. It shows the visual representation of the results obtained in Table 3 where it is seen clearly that the best performance is shown by Puk-kernel model, the observed discharge is well captured by this model throughout the time period. Underestimation of high flows can be observed from the figure for this model comparison of observed streamflow. Discharge of more than $\sim 1500 \text{ m}^3/\text{s}$ is not well captured in this model. Low flows are very captured in the model, which shows that baseflow is well simulated. However, when coming to the other models in the scenario 1, RBF-

Kernel is the second best model, where it highly underestimates the peak flows and simulates the low flow quite well. This shows the incapability of all the models of capturing the high flows.

Table 4. Statistical indicators and error measures between observed and simulated discharge of four different models utilized in this study for scenario 2.

Model	Dataset	r	MAE m ³ /s	RMSE m ³ /s	RAE (%)	RRSE (%)
Poly Kernel	Training	0.878	152.299	196.923	72.615	56.426
	Testing	0.907	212.693	251.529	85.316	66.007
RBF-Kernel	Training	0.853	88.348	202.819	42.124	58.116
	Testing	0.885	125.204	230.928	50.222	60.601
Norm. Poly Kernel	Training	0.813	123.507	205.506	58.887	58.886
	Testing	0.553	195.942	319.291	78.597	83.789
Puk-Kernel	Training	0.969	34.190	89.194	16.301	25.558
	Testing	0.974	42.912	90.459	17.213	23.738

Fig. 5. Reproduction of observed and simulated discharge for four different models as a function of time a) Poly Kernel, b) RBF-Kernel, c) Norm. Poly Kernel, and d) PUK-Kernel for scenario 2

Table 5. Statistical indicators and error measures between observed and simulated discharge of four different models utilized in this study for scenario 3.

Model	Dataset	r	MAE m ³ /s	RMSE m ³ /s	RAE (%)	RRSE (%)
Poly Kernel	Training	0.879	147.733	192.161	70.438	55.060
	Testing	0.892	206.243	247.396	82.729	64.92
RBF-Kernel	Training	0.847	89.314	204.081	42.584	58.478
	Testing	0.868	121.667	229.473	48.803	60.219
Norm. Poly Kernel	Training	0.823	119.171	200.937	56.820	57.580
	Testing	0.5806	192.2897	312.6158	77.1316	82.0377
PUK-Kernel	Training	0.970	33.407	88.346	15.928	25.315

	Testing	0.971	42.635	94.979	17.102	24.925
--	---------	-------	--------	--------	--------	--------

Fig. 6. Reproduction of observed and simulated discharge for four different models as a function of time a) Poly Kernel, b) RBF-Kernel, c) Norm. Poly Kernel, and d) PUK-Kernel for scenario 3

366

367 In the scenario 2, similar to the previous case the best results obtained are from the PUK-Kernel
 368 model with $r = 0.97$, $MAE = 42.91 \text{ m}^3/\text{s}$, $RMSE = 90.45 \text{ m}^3/\text{s}$, $RAE = 17.21 \%$, $RRSE = 23.73 \%$
 369 in the testing period. Whereas, the least performance in the testing period shown by Norm. Poly
 370 Kernel with $r = 0.55$, $MAE = 195.94 \text{ m}^3/\text{s}$, $RMSE = 319.29 \text{ m}^3/\text{s}$, $RAE = 78.59 \%$, $RRSE = 83.78$
 371 $\%$. Rest of the models performed in between these model performances as shown in the Table 4
 372 with detailed error measures. The clear visualization of these statistical measures can be further
 373 illustrated from the Fig.5 where PUK-Kernel simulated the discharge against observed quite well
 374 with underestimation in the peak flow. From Figs. 6-8, it can be observed that a similar trend of
 375 the discharge is obtained from all the five models with PUK-Kernel performing the best model.
 376 The r value of PUK-Kernel from scenario 3 to 5 is ~ 0.97 . While for all other models, it ranged
 377 from 0.48 (Norm. Poly Kernel) to 0.92 (Poly Kernel).

378 **Table 6.** Statistical indicators and error measures between observed and simulated discharge of
 379 four different models utilized in this study for scenario 4.

Model	Dataset	r	MAE m^3/s	RMSE m^3/s	RAE (%)	RRSE (%)
Poly Kernel	Training	0.881	142.093	186.954	67.749	53.570
	Testing	0.885	200.304	242.435	80.346	63.621
RBF-Kernel	Training	0.845	89.769	204.491	42.801	58.595
	Testing	0.858	120.684	229.946	48.409	60.343
Norm. Poly Kernel	Training	0.828	115.793	198.245	55.209	56.805
	Testing	0.611	187.879	305.075	75.363	80.059
PUK-Kernel	Training	0.971	32.958	86.645	15.714	24.827
	Testing	0.970	43.652	97.090	17.510	25.479

380

Fig.7. Reproduction of observed and simulated discharge for four different models as a function of time a) Poly Kernel, b) RBF-Kernel, c) Norm. Poly Kernel, and d) PUK-Kernel for scenario 4

387 **Table 7.** Statistical indicators and error measures between observed and simulated discharge of
388 four different models utilized in this study for scenario 5.

Model	Dataset	r	MAE m ³ /s	RMSE m ³ /s	RAE (%)	RRSE (%)
Poly Kernel	Training	0.884	137.426	182.408	65.524	52.267
	Testing	0.883	195.146	235.941	78.277	61.916
RBF-Kernel	Training	0.845	90.054	204.588	42.937	58.623
	Testing	0.852	120.485	230.626	48.329	60.522
Norm. Poly Kernel	Training	0.831	113.224	197.023	53.984	56.455
	Testing	0.634	186.242	298.323	74.706	78.287
PUK-Kernel	Training	0.969	33.644	89.934	16.041	25.770
	Testing	0.971	44.208	97.138	17.733	25.491

389

390

Fig. 8. Reproduction of observed and simulated discharge for four different models as a function of time a) Poly Kernel, b) RBF-Kernel, c) Norm. Poly Kernel, and d) PUK-Kernel for scenario 5. The results in all scenarios were also evaluated through Taylor diagram as shown in Fig. 9. The Taylor diagram represents three indices namely, correlation coefficient, standard deviation and root mean square difference (RMSD) (Taylor, 2001). Based on Taylor diagram, it is clear that the PUK kernel model was found to be the superior while Normalized Poly Kernel model were observed worst among all scenarios.

Fig. 9. Taylor diagram for developed models during testing periods in all scenarios

Previous studies on the stream discharge relationship have demonstrated that PUK-Kernel model performs better than other models utilized here (Singh et al., 2019; Sattari et al., 2021). In a recent study by Khosravi et al. (2019) have compared the different methods based on the kernel and standalone tree based algorithms and found that the prediction power of kernel based functions are higher than non-kernel based functions. The findings of the study are corroborated with our results and support firmly the results with the error measures and graphical indicators obtained through model simulations (Kumari and Srivastava, 2020; Elbeltagi et al., 2021). Sattari et al., 2021 showed that poly kernel function provides the better results than the RBF kernel also they found that different scenarios yielded the same results of poly kernel functions. These findings are in line with our study where in all the five scenarios our best model does not change depending on the scenarios selected. The similar results were also obtained in other previous studies (Srivastava et al., 2017; 2018). Rasouli et al., 2011 demonstrated that nonlinear models outperformed the multiple linear regression models and Bayesian neural network slightly outperformed the linear modes in the simulation of lead times of streamflow. In addition, the

least performance shown in this study by normalized poly kernel is supported with the previous literature (Wang et al., 2009). Further, Zhu et al., 2019 also showed that seasonal streamflow is better simulated with kernel GPR method that well relates with our study. In another carried out by Adamala and Srivastava, 2018 it was found that the kernel based methods are one of the best alternatives to understand the rainfall-runoff relationship in the catchment. In particular to this climatology, the discharge estimation under predicts the observed streamflows which may due to the to the overprediction of model simulated water balance losses, such as base flow and ET (Paul et al., 2018; Srivastava et al., 2020; Elbeltagi et al., 2020).

4. Conclusions

The present study has examined four techniques of Gaussian processes regression (GPR): Polynomial Kernel, Radial Basis Function Kernel, Normalized Polynomial Kernel and PUK Kernel, to capture the daily discharge of Govindpur site located at the Burhabalang river in the India. The proposed models were trained and tested by using 10-year daily river stage and discharge data is measured from 2004 to 2013. The datasets were divided from 2004 to 2010 for models training process and from 2011 to October 2013 for testing. These models' precision was optimized by considering five different scenarios based on various combinations of time-lags of stage and discharge values up to 5-day. The performance of developed models was evaluated by r , RMSE, RAE, RRSE and visual inspection like line diagram and scatter plot. The results showed that the PUK kernel model outperforms all the other proposed models in each of the five examined scenarios during both training and testing phases. In contrast, the Normalized poly kernel model is the lowest performance model in all scenarios for predicting the river discharge.

Thus, it can be concluded that the PUK kernel model could be used as a reliable model for river discharge prediction and enhancing the flood management system.

Authors' contributions

Ahmed Elbeltagi-Manish Kumar: Conceptualization, Formal analysis, Writing - original draft, Writing - review & editing. Ahmed Elbeltagi: data processing and modeling; Manish Kumar: preparing tables and figures; Ankur Srivastava: results writing; Anuradha Kumari: data collection; Rawshan Ali, Chaitanya Pande : writing methods and introduction; Abu Reza Md. Towfiqul Islam, Daniel Prakash Kushwaha: writing discussion and editing. All coauthors: Writing- review & editing. All authors read and approved the final manuscript.

Ethics approval and consent to participate: Not applicable

Competing interests: The authors declare that they have no competing interests

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Consent for publication: Not applicable

Availability of data and materials: The datasets generated during and/or analyzed during the current study are available from corresponding authors on reasonable request.

References

- Adamala, S., Srivastava, A. 2018. Comparative evaluation of daily evapotranspiration using artificial neural network and variable infiltration capacity models. *Agricultural Engineering International: CIGR Journal*, J 20(1):32–39
- Adams, R. and Stegle, O. 2008. Gaussian process product models for nonparametric nonstationarity. In *Proceedings of the 25th international conference on Machine learning*. ACM.

461 Alvarez, M. and Lawrence, N. 2008. Sparse convolved Gaussian processes for multi-output
 462 regression. In NIPS.

463 Alvarez, M. and Lawrence, N. 2011. Computationally efficient convolved multiple output
 464 gaussian processes. *Journal of Machine Learning Research*, 12:1425-1466.

465 Bajirao, T.S., Kumar, P., Kumar, M., Elbeltagi, A. and Kuriqi, A., 2021. Superiority of Hybrid
 466 Soft Computing Models in Daily Suspended Sediment Estimation in Highly Dynamic
 467 Rivers. *Sustainability*, 13(2), p.542.

468 Bonilla, E., Chai, K., and Williams, C. 2008. Multi-task Gaussian process prediction. In NIPS.

469 Boyle, P. and Frean, M. 2004. Dependent Gaussian processes. In NIPS.

470 Cortes, C., Vapnik, V. 1995. Support-Vector Networks. *Machine Learning*, Kluwer Academic

471 Dawid, A. 1973. Posterior expectations for large observations. *Biometrika*, 60(3):664.

472 De Finetti, B. 1956. The Bayesian approach to the rejection of outliers. In *Proceedings of the*
 473 *Fourth Berkeley Symposium on Mathematical Statistics and Probability: held at the Statistical*
 474 *Laboratory, University of California, June 30-July 30, 1960*, page 199. University of California
 475 Press.

476 Elbeltagi, A., Aslam, M. R., Malik, A., Mehdinejadani, B., Srivastava, A., Bhatia, A. S., & Deng,
 477 J. 2020. The impact of climate changes on the water footprint of wheat and maize production in
 478 the Nile Delta, Egypt. *Science of the Total Environment*, 743, 140770.
 479 <https://doi.org/10.1016/j.scitotenv.2020.140770>

480 Elbeltagi, A., Kumari, N., Dharpure, J. K., Mokhtar, A., Alsafadi, K., Kumar, M., ... & Kuriqi, A.
 481 2021. Prediction of Combined Terrestrial Evapotranspiration Index (CTEI) Over Large River
 482 Basin Based on Machine Learning Approaches. *Water*, 13(4), 547.

483 Fox, E. and Dunson, D. 2011. Bayesian nonparametric covariance regression. Arxiv
484 preprint arXiv:1101.2017.

485 Gibbs, M. 1997. Bayesian Gaussian Process for Regression and Classification. PhD thesis, Dept.
486 of Physics, University of Cambridge.

487 Goldberg, P. W., Williams, C. K., and Bishop, C. M. 1998. Regression with input-dependent
488 noise: A Gaussian process treatment. In NIPS.

489 Goldberg, Y., Elhadad, M. 2008. SplitSVM: Fast, Space-Efficient, non-Heuristic, Polynomial
490 Kernel Computation for NLP Applications. Proc. ACL-08: HLT.

491 Goovaerts, P. 1997. Geostatistics for natural resources evaluation. Oxford University Press, USA.

492 Inamdar, N.M., Ehrlich, K.C., Ehrlich, M., et al., 2004. Data mining in bioinformatics using Weka.
493 Bioinformatics 20, 2479–2481.

494 Keerthi, S. S., Lin, C.J. 2003. Asymptotic behaviors of support vector machines with Gaussian
495 kernel. Neural Computation, 15(7):1667–1689.

496 Kersting, K., Plagemann, C., Pfaff, P., and Burgard, W. 2007. Most likely heteroscedastic Gaussian
497 process regression. In Proceedings of the 24th international conference on Machine learning,
498 pages 393-400. ACM.

499 Khosravi, K., Shahabi, H., Pham, B. T., Adamowski, J., Shirzadi, A., Pradhan, B., ... & Prakash, I.
500 2019. A comparative assessment of flood susceptibility modeling using multi-criteria decision-
501 making analysis and machine learning methods. Journal of Hydrology, 573, 311-323.

502 Kumar, M. 2020. Estimation of Daily Stage–Discharge Relationship by Using Data-Driven
503 Techniques of a Perennial River, India." Sustainability 12.19: 7877.

504 Kumari, A., Kumar, M., Kumar, A., Amitabh, A. 2020. Daily Gauge-Discharge Estimation using
505 ANN and Wavelet-ANN Models for Muri Station, Jharkhand. *Indian Journal of Ecology* 47(3):
506 645-649.

507 Kumari, N., Srivastava, A. 2020. An Approach for Estimation of Evapotranspiration by
508 Standardizing Parsimonious Method. *Agric Res* 9, 301–309. [https://doi.org/10.1007/s40003-019-](https://doi.org/10.1007/s40003-019-00441-7)
509 [00441-7](https://doi.org/10.1007/s40003-019-00441-7)

510 Kuss, M. and Rasmussen, C. 2005. Assessing approximate inference for binary Gaussian process
511 classification. *The Journal of Machine Learning Research*, 6:1679-1704.

512 Lazaro-Gredilla, M. and Titsias, M. 2011. Variational heteroscedastic gaussian process regression.

513 Lin, H.T., Lin. C.J. 2003. A study on sigmoid kernels for SVM and the training of non-PSD
514 kernels by SMO-type methods. Technical report, Department of Computer Science, National
515 Taiwan University.

516 Maass, W. 2000. On the Computational Power of Winner-Take-All. *Neural Computation*. 12,
517 2519-2535.

518 Manfreda, S.; Pizarro, A.; Moramarco, T.; Cimorelli, L.; Pianese, D.; Barbetta, S. 2020. Potential
519 advantages of flow-area rating curves compared to classic stage-discharge-relations. *J. Hydrol.*
520 585, 124752, doi:10.1016/j.jhydrol.2020.124752.

521 Markonis, Y., Koutsoyiannis, D. 2015. Scale-dependence of persistence in precipitation records.
522 *Nature Climate Change* 6, 399–401.

523 Mohanty, P.K.; Mohanty, L.P.; Khatua, K.K. 2019. Discharge estimation in wide meandering
524 compound channels. *ISH J. Hydraul. Eng.* 25, 1–15.

525 Neal, R. 1996. Bayesian learning for neural networks. Springer Verlag.

526 O'Hagan, A. 1979. On outlier rejection phenomena in Bayes inference. Journal of the Royal
527 Statistical Society. Series B (Methodological), pages 358-367.

528 Osborne, M., Roberts, S., Rogers, A., Ramchurn, S., and Jennings, N. 2008. Towards real-time
529 information processing of sensor network data using computationally efficient multi-output
530 gaussian processes. In Proceedings of the 7th international conference on Information processing
531 in sensor networks, pages 109-120. IEEE Computer Society.

532 Paul, P. K., Kumari, N., Panigrahi, N., Mishra, A., & Singh, R. 2018. Implementation of cell-to-
533 cell routing scheme in a large scale conceptual hydrological model. Environmental Modelling &
534 Software, 101, 23-33. <https://doi.org/10.1016/j.envsoft.2017.12.003>

535 Rasmussen, C. E. 1996. Evaluation of Gaussian Processes and Other Methods for Non-linear
536 Regression. PhD thesis, Dept. of Computer Science, University of Toronto.

537 Rasmussen, C. E. and Williams, C. K. 2006. Gaussian processes for Machine Learning. The MIT
538 Press.

539 Rasmussen, C., Williams, C. 2012. Gaussian Processes for Machine Learning. MIT Press, Boston,
540 Mass pp. 69–106

541 Rasouli, K., Hsieh, W. W., & Cannon, A. J. 2012. Daily streamflow forecasting by machine
542 learning methods with weather and climate inputs. Journal of Hydrology, 414, 284-293.

543 Sattari, M. T., Feizi, H., Colak, M. S., Ozturk, A., Apaydin, H., & Ozturk, F. 2020. Estimation of
544 sodium adsorption ratio in a river with kernel-based and decision-tree models. Environmental
545 Monitoring and Assessment, 192(9), 1-13.

546 Schmidt, A. and O'Hagan, A. 2003. Bayesian inference for non-stationary spatial covariance
547 structure via spatial deformations. Journal of the Royal Statistical Society: Series B (Statistical
548 Methodology), 65(3):743-758.

549 Scholkopf, B. 1997. Support Vector Learning. R. Oldenburg Verlag, Munich.

550 Scholkopf, B., Burges, C.J.C., Smola. A.J. 1999. Advances in Kernel Methods. The MIT Press,
551 MA.

552 Singh, B., Sihag, P. & Deswal, S. Modelling of the impact of water quality on the infiltration rate
553 of the soil. Appl Water Sci 9, 15 2019. <https://doi.org/10.1007/s13201-019-0892-1>

554 Srivastava, A., Kumari, N. & Maza, M. 2020. Hydrological Response to Agricultural Land Use
555 Heterogeneity Using Variable Infiltration Capacity Model. Water Resour Manage 34, 3779–
556 3794. <https://doi.org/10.1007/s11269-020-02630-4>

557 Srivastava, A., Sahoo, B., Raghuwanshi, N. S., & Chatterjee, C. 2018. Modelling the dynamics of
558 evapotranspiration using Variable Infiltration Capacity model and regionally calibrated
559 Hargreaves approach. Irrig Sci 36, 289–300 (2018). <https://doi.org/10.1007/s00271-018-0583-y>

560 Srivastava, A., Sahoo, B., Raghuwanshi, N. S., & Singh, R. 2017. Evaluation of variable-
561 infiltration capacity model and MODIS-terra satellite-derived grid-scale evapotranspiration
562 estimates in a River Basin with Tropical Monsoon-Type climatology. Journal of Irrigation and
563 Drainage Engineering, 143(8), 04017028. [https://doi.org/10.1061/\(ASCE\)IR.1943-4774.0001199](https://doi.org/10.1061/(ASCE)IR.1943-4774.0001199)

564 Sun, A. Y., Wang, D., Xu, X. 2014. Monthly streamflow forecasting using Gaussian Process
565 Regression. Journal of Hydrology 511, 72–81.

566 Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single
567 diagram. *Journal of Geophysical Research: Atmospheres*, 106(D7), pp.7183-7192.

568 Teh, Y., Seeger, M., and Jordan, M. 2005. Semiparametric latent factor models. In Workshop on
569 Artificial Intelligence and Statistics, volume 10.

570 Turner, R. and Sahani, M. 2008. Modeling natural sounds with modulation cascade processes.
571 Advances in neural information processing systems, 21.

572 Turner, R. E. 2010. Statistical Models for Natural Sounds. PhD thesis, University College
573 London.

574 Uestuen, B., Melssen, W.J., Buydens, L.M.C., 2006. Facilitating the application of support vector
575 regression by using a universal Pearson-function based kernel. *Chemometrics and Intelligent*
576 *Laboratory Systems* 81, 29–40.

577 Vanhatalo, J., Jylanki, P., and Vehtari, A. 2009. Gaussian process regression with Student-t
578 likelihood. In *NIPS*.

579 Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, NY.

580 Wang, W. C., Chau, K. W., Cheng, C. T., & Qiu, L. 2009. A comparison of performance of
581 several artificial intelligence methods for forecasting monthly discharge time series. *Journal of*
582 *hydrology*, 374(3-4), 294-306.

583 Wilson, A. G. and Ghahramani, Z. 2010a. Copula processes. In *NIPS*.

584 Wilson, A. G. and Ghahramani, Z. 2010b. Generalised Wishart Processes. *Arxiv*
585 *preprint arXiv:1101.0240*

586 Zhu, S., Luo, X., Xu, Z., & Ye, L. 2019. Seasonal streamflow forecasts using mixture-kernel GPR
587 and advanced methods of input variable selection. *Hydrology Research*, 50(1), 200-214.

588

589

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ConflictofInterest.doc](#)
- [GraphicalAbstract.docx](#)
- [Highlights.docx](#)