VIETNAM JOURNAL OF HYDROMETEOROLOGY

Check for updates

Research Article

A comparative analysis of regression equations for rating curve development at a gauging station in Da river, Northern Vietnam

Minh Dang Tran Duc[1], Huy Dao Ba[1], Quynh Hoang Diem[1], Tinh Nguyen Thi[2], Hanh Nguyen Duc[1], Vinh Tran Ngoc[3], Giang Nguyen Tien[1*]

[1] Faculty of Hydrology, Meteorology and Oceanography, University of Science, Vietnam National University, Hanoi. Add: 334 Nguyen Trai Street, Thanh Xuan District, Hanoi, Viet Nam; dangtranducminh_t65@hus.edu.vn; daobahuy_t66@hus.edu.vn; diemquynhoang918@gmail.com; nguyenduchanh@hus.edu.vn; nguyentiengiang@hus.edu.vn

[2] Center for hydro–meteorological observation, Viet Nam Meteorological and Hydrological Administration. Add: 8 Phao Dai Lang, Lang Thuong, Dong Da, Ha Noi, Viet Nam; tinh.nt.198@gmail.com

[3] Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI, USA; vinhtn@umich.edu

*Corresponding author: nguyentiengiang@hus.edu.vn; Tel: +84–912800896

Received: 5 February 2023; Accepted: 23 March 2023; Published: 25 March 2023

Abstract: Constructing rating curves at hydrological stations is of tremendous significance for water resources management, yet in Vietnam, it has not been given the adequate attention it deserves. In this study, eight traditional regression equations representing the linear and non–linear correlation between gauging discharge and water level (stage) at PoLech station in Da river, were evaluated with the aim of determining the most suitable equations for discharge interpolation and high flow extrapolation. A straightforward segmentation technique was proposed to simplify the automatic piecewise regression. The results revealed that: i) Second–order polynomial regression equations (in which stage is independent variable and either Q or $Q^{1/2}$ is dependent variable) proved to be the most efficient for discharge interpolation, when automatic piecewise regression was applied; ii) The linear regression equation illustrating relationship between square root of discharge and stage performed the best for high flow extrapolation; iii) The amalgamated rating curve, which was formed by utilizing all–years rating data, could be used for each year interpolation with care and additional research is required in relation to its accuracy. The potential of being able to generate continual discharge estimations at a low cost and with relatively uncomplicated calibration methods is expansive. This approach has the potential to encourage researchers, aquatic ecosystem stewards, water quality monitors, or appraisers of upstream withdrawals to start gauging river discharge on a more regular basis from an operational standpoint.

Keywords: Regression; Rating curve; Da river; Interpolation; Extrapolation of discharge.

1. Introduction

The effective management of water resources necessitates the utilization of hydrologic variables such as precipitation, run–off, or discharge in streams [1–5]. The amount of water circulating in streams can significantly differ in both temporal and spatial terms, which is primarily attributed to the variations in duration, frequency, intensity, and extent of precipitation as well as the characteristics of the catchment area [5–10]. Knowledge of the

flow of rivers and their variability is an essential part of the assessment and management of surface water resources. To conduct reservoir designs, flood frequency studies, flood inundation modeling, design of flood protection and warning systems, water supply engineering, drought studies, geomorphologic studies, etc., the records of discharge measurement need to be obtained from the river [11, 12]. However, this process is often costly, laborious and can be difficult to conduct in the case of extreme floods [11, 13].

By constantly monitoring the water level, it is possible to accurately calculate river discharge through the application of a rating curve, which is a correlation between discharge and water level [14–17]. When the hydraulic properties of a river remain constant and the stage–discharge relationship is not influenced by unsteadiness, it is relatively easy to develop a reliable single–valued rating curve (a one–to–one relationship between the stage and the discharge) [17–19]. There have been many studies on building stage–discharge relationships, which followed two main approaches: the physically based approach and the data–driven approach. In a physically based approach, Manning's equation – a widely accepted and extensively utilized physical equation is typically incorporated in 1D, 2D, and 3D hydrodynamic models to quantify the interrelationship between discharge and hydraulic head [12, 20, 21]. This approach requires accurate information regarding the topography of the channel and boundary conditions of the flow. The data–driven approach is based on a relationship (linear or nonlinear) between discharge and stage and possibly other related factors such as velocity, slope, bed roughness, etc. In this paper, the data–driven approach is the focus, so the following will refer to this approach in more detail.

Following the data–driven approach, there are three groups of methods, namely: graphical [14, 22]; regression [15, 23–26]; and machine learning [27–28]. According to [5], many researchers use Machine Learning algorithms such as Artificial Neural Network (ANN), Support Vector Machine (SVM), MT, Takagi–Sugeno (TS) fuzzy inference, Genetic Algorithm (GA), Generalized Reduced Gradient (GRG) to derive discharges from other measured variables. However, these methods often require additional data (stage and/or discharge) from other stations upstream of the station under study. Therefore, in this paper, the traditional methods are used to ensure that it is possible to build rating curves that can serve to calculate the discharge at a station from only the gauged stage of that station.

The efficacy of conventional regression techniques has been established through a plethora of preceding examinations and is frequently employed. Essentially, the rating curves are calculated through a regression process with the use of measurements of stages (water surface elevation above the mean sea level) and the corresponding measured discharges, and sometimes from velocity distribution, bed roughness and friction slope [18]. [29] developed two methods for automatic computation using least–squares approximation, one based on polynomials and the other on piecewise–continuous splines. Both methods were found to work well and once the parameters for a gauging station have been determined, rating data can be processed automatically. According to [29], whereas it is sometimes a convenient approximation to the relationship Q–H over the whole range of data, in general it is an over–simplification of the real hydraulics at many gauging stations. The more general representation of Q–H relationship by a polynomial of higher degree M has been in the background for some time [1, 22, 30–32] used it successfully with just M = 3, and in general, most of all studies show that with M greater than 3, the results are not good. [14] suggested writing the polynomial for Q raised to the power (Q). [33] used the polynomial approximation methods to obtain 622 rating curves from 171 Australian Bureau of Meteorology Hydrologic Reference Stations. They found that the methods worked well except for about 0.5% of the stations, where there was difficulty approximating the low–flow data. Despite the widespread use of regression equations for flow calculation

worldwide, very few studies have been conducted to explore the suitability of such equations for Vietnamese locations that only record water level measurements. In particular, [34] suggested using the Spline regression function of third order to develop rating curves for 23 stations across 19 rivers in Vietnam. [35] proposed the use of linear and second–order nonlinear regression for rating curve development based on data from the Ha Bang hydro–station on the Ky Lo river from 2013 to 2020. A commonality among these few studies is that the authors attempted to propose a single rating curve based on a long sequence of data. However, the water level–discharge relationship often changes over time due to changes in topography/riverbed morphology, particularly in rivers affected by large reservoir systems such as the Da River. As a result, there is an urgent need to investigate the potential and effectiveness of these techniques in this context.

Specifically, in Vietnam, the discharge compilation/processing is typically conducted in accordance with the stringent provisions of the regulation of TCVN 12636–15:2021 [36]. Generally, hydrological stations in Vietnam have been relying on manual direct measurement of flow using flowmeters, floats, and acoustic doppler current profiler devices, with the subsequent manual correction achieved through a series of steps. This process involves the drawing of a cross–section of a river, investigating of the cross–relationship of factors such as discharge, water level, cross–sectional area, flow velocity and the analysis of the relationship between these factors, and selecting of the most suitable processing method. For the later task, the calculation of rating curves is usually done manually (drawn on technical papers) or using HydPro1.0 software. In this software, the following equations can be used to build a steady–flow rating curve: exponential function, polynomial of order 2 (parabola), Spline regression function of order 3; Q as a polynomial second degree of H. There have been almost no domestic studies suggesting which function is the best for rating curve construction or evaluating the possibility of rating curve extrapolation. Meanwhile extrapolation is one of the important roles of using rating curves. Therefore, this paper is focused on answering the following three research questions:

1. Which regression equation is the best for interpolating discharge given pairs of gauging Q and H for an individual year of records?

2. Is it possible to apply a regression equation using multi–year gauging data to interpolate Q for each year in a stable station?

3. Which regression equation is the best for extrapolating high discharges given pairs of gauging Q and H for an individual year of records?

Compared with the current processing approach, the utilization of regression–based approaches not only ensure the precision of flow estimation, but also boasts superior performance in terms of efficiency, capability, and transparency. Specifically, with the proposed equation, the flow correction can be accomplished expeditiously in comparison to the existing manual labor. Additionally, if the data is refreshed automatically from the automated measuring stations, the provision of real–time data can be done with considerable ease. Furthermore, with new data on both discharge and water level being updated, the equations can automatically refresh and re–optimize the coefficients rapidly through optimization algorithms. Lastly, with the proposed equations being made available to the public through this study, their application and utilization will be more widespread among numerous target groups such as administrators or researchers. In contrast, the current data processing and editing in Vietnam is only conducted internally and not made public to the relevant units. The advantage of the suggested approach is particularly prominent for stations with unique locations (e.g., PoLech station, Da river, which is the primary focus of this work) located on transnational rivers and situated at the upstream position of the largest reservoir system in Vietnam.

2. Materials and Methods

2.1. Study site

Da river originates from Yunnan province, China. The whole basin stretches from 20º40'N – 25º00'N and 100º22' – 105º24'W [37]. Its total length is approximately 1010 km with a catchment area of around 52900 km^2 (49% of area belongs to China) [38]. In Vietnam, Da river flows through Lai Chau, Dien Bien, Son La, Hoa Binh, Phu Tho provinces. The river discharge is large, which accounted for 31% of water supply for the Red river basin. Da river is in a tropical monsoon climate. The rainfall distributed unevenly in time and space. The annual rainfall in the Da River basin in the period 2016–2021 ranges from 1,300 mm to over 2,600 mm [39]. The rainy season starts from May and lasts in October, with rainfall accounting for 85 to 90% of the total annual rainfall [37]. The flow regime of the Da River is heavily influenced by rainfall, with the flood season occurring from June to October and reaching a peak in July and August. The dry season lasts from November to May of the following year.

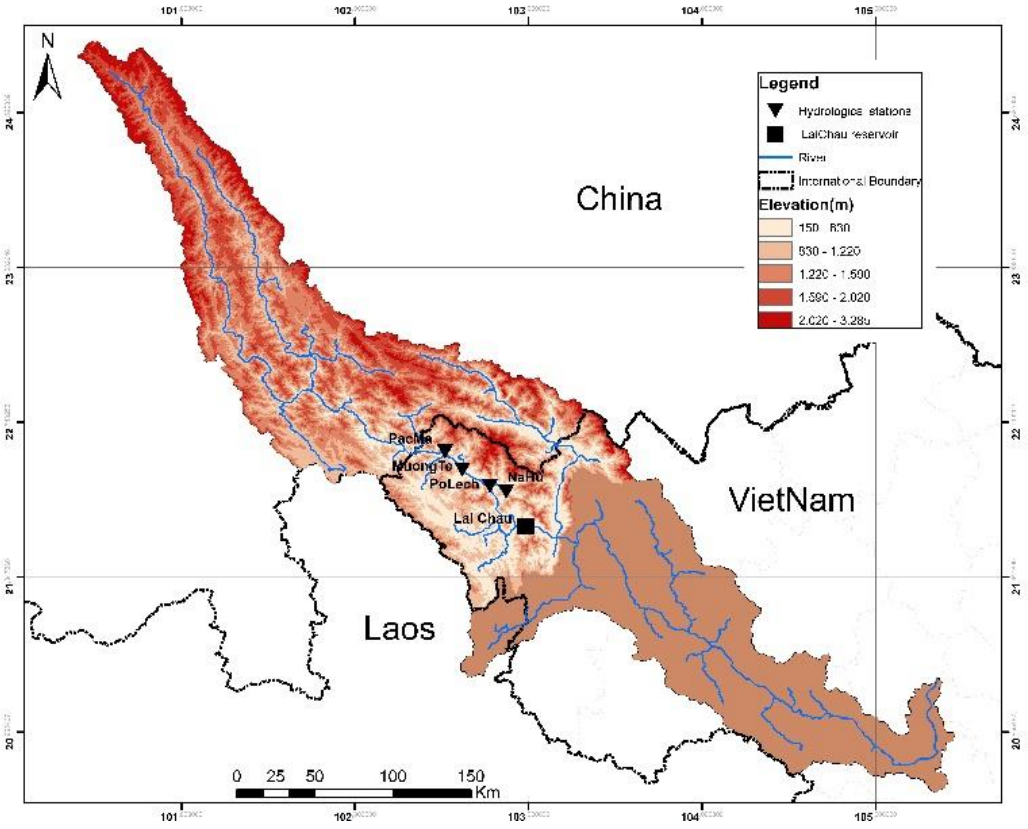

Figure 1. Da river basin and PoLech station.

The PoLech station is located on Da river in Muong Te district, Lai Chau province, Vietnam (Figure 1). Catchment area at PoLech station is 26270 km^2. PoLech is hydrological station level 1, which was established in 2003 to serve the Son La hydropower plant design and construction and continued to measure hydrological data from 2007 to 2011 for designing the Lai Chau hydropower dam. The station is located on the right bank of the Da River, 12 km downstream of the Muong Te level 3 hydrological station. The station is located on a quite straight reach of the river, with an average width of around 100 m. The gauging discharges and stages as well as recorded stages were obtained at the same cross–section. Currently, the correction of discharge at Polech station is carried out according to the standard regulation TCVN 12636–15:2021. Figure 2 illustrates the cross–section of the station in 2006 and there were no significant changes over the years. The left bank is a steep mountain slope with stable geological and topographical conditions, while

the right bank has a rocky bottom at the lower part and sand and sediment at the upper part. The highest stage measured at the station is 275.40 m. Therefore, the cross–section controls the highest water level and there is no clear floodplain delineation.

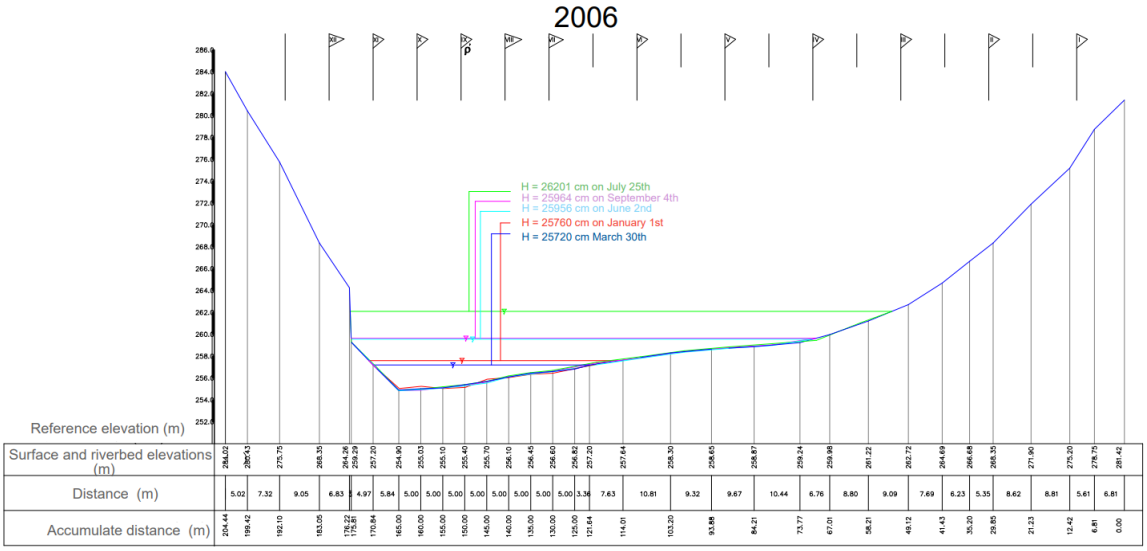

Figure 2. River cross–section at PoLech hydrometric station in 2006.

2.2. Data collection

Gauging stages and discharges in the period 2005–2011 of PoLech station are used in this study. The stage data measured at the same time as the observed discharge is called the gauging stage. The recorded stage is the stage measured frequently and without discharge measurement at the same time. Table 1 shows the number of gaugings (number of gauged Q, H pairs), the maximum and minimum values of gauging stage and discharge (H^{g}_{max}, H^{g}_{min}) and the recorded stage (H^{re}_{max}, H^{re}_{min}) of each year. Note that the gauging data or recorded data obtained from in situ measurement is different from the rating data, which is simulated from a regression equation.

Table 1. Statistics on stage and discharge data used in this study.

Year	2005	2006	2007	2008	2009	2010	2011	2005–2011
No. of gaugings	79	62	55	40	60	49	53	398
H^{re}_{max} (cm)	26790	26870	27070	26614	26700	26489	26455	27070
H^{re}_{min} (cm)	25679	25709	25671	25688	25704	25666	25689	25666
H^{g}_{max} (cm)	26784	27251	27113	26590	26637	26487	26403	27251
H^{g}_{min} (cm)	25681	25720	25678	25688	25705	25671	25700	25671
Q^{g}_{max} (m^3/s)	3810	6460	5570	2950	3050	2458	2068	6460
Q^{g}_{min} (m^3/s)	56.8	111	42.7	65.9	84.9	47.4	81.7	42.7

3. Methods

3.1. Data preprocessing

The raw hydrologic data often contains some uncertainty due to the measurement or data processing. The uncertainty in the data may lead to incorrect reflections in calculating, analyzing hydrologic data, and therefore in making decisions. Thus, outlier detection and removal in the data set are critical for improving calculation quality. There are several methods to identify the outliers, such as standard deviation, the interquartile range [40]. Because the gauging data is not continuous and does not have a normal distribution, this

research used the interquartile range theory to identify the outliers. The Interquartile range is the range of values between ¼ and ¾ positions of the ascending order data. The first and the third quartile are the 25th (Q_1) and 75th percentiles (Q_3) of the data set, respectively. The thresholds for investigating the outliers are defined as follows:

Lower threshold: $L = Q_1 - 1.5(Q_3 - Q_1)$

Upper threshold: $U = Q_3 + 1.5(Q_3 - Q_1)$

where $(Q_3 - Q_1)$ is the interquartile range (IQR).

The outliers are the points having smaller values than Lower threshold or greater values than Upper threshold.

3.2. Determining the stability of the Stage – Discharge relationship

According to the Vietnamese national standard [36], a stable Stage–Discharge relationship is represented by a smooth rating curve that has a unique discharge corresponding to each stage and passes through the center of groups of data. In addition, the stable rating curve must comply with the following requirements: (1) having a balanced number of points on either side of the curve; (2) balancing the negative and positive error; (3) the error (σ) of the rating curve calculated in section 3.4 should be smaller than five percent. In addition, to determine the stability of the stage–discharge relationship, [18] suggested: (4) plotting the gauging data points on both the up and down curve of water level to identify the loop relationship between stage and discharge or any other relationship; (5) comparing the gauging and rating data from 2 to 5 years to investigate any significant changes; and (6) plotting the gauging and recorded data during the flood and dry seasons to observe if seasonal factors such as plant growth and/or sedimentation have any influence. In this study, graphs and performance metrics representing these criteria were used to determine the stability of the stage–discharge relationship.

3.3. Seelcting regression equations

3.3.1 Traditional regression equations

A regression equation is used to estimate the relationship between the dependent variable (an unknown variable) and the independent variable (a known variable). Discharge is considered a dependent variable with a short data series that is difficult to measure continuously; water level (stage) is an independent variable with longer data and is easier to measure. Therefore, regression is a method for calculating discharge that is based on establishing the relationship between stage and discharge. Linear regression and non–linear regression are two popular methods. Linear regression equation represents the relationship between variables as a straight line. Non–linear regression equation represents the relationship between variables as a curve. Non–linear regression equations could be polynomial or power. Table 2 shows the types and the corresponding mathematical expressions of regression equations used in this study.

Table 2. Traditional regression equations.

Types of equations		**Regression Equations**	
Linear		$Q = a + H.b$	(1)
		$Q = a + (H - Ho).b$	(2)
	Logarithm	$\text{Log}\, Q = b.\log(H - Ho) + \text{loga}$	(3)
		$Q^v = a + H.b$	(4)
Non– linear	Power equation	$Q = a.(H - Ho)^b$	(5)
	Polynomial second order	$Q = a + b.H + c.H^2$	(6)
		$Q = a + b.(H - Ho) + c.(H - Ho)^2$	(7)
		$Q^v = a + b.H + c.H^2$	(8)

Depending on the order of the dependent variable, the relationship types will reflect the results differently. According to [12], using ν = ½ gives a good approximation result that is useful for calculation, and stabilizes the variance, especially for small flows. Therefore, in this study, ν = ½ was used in equations (4) and (8).

3.3.2. Piecewise regression

It is necessary to break the regression line into multiple segments (piecewise) when: (1) The river cross–section is unsteady over time (sedimentation, seasonal plant growth or other changes affecting hydraulic characteristics of river); (2) Although the river cross–section remains stable over time, variations in the Q–H relationship due to changes in both the cross–sectional shape and the riverbank. The segmentation of a stable river cross–section is oftentimes based on the shape of cross–section or the rating curve. For the cross–sections with unclear stage differentiation (i.e., no clear separation between riverbed and riverbank), breaking the regression line is difficult and inactive. In this research, a simple and automatic method was proposed to break the regression curve into 4 segments: i) low flows; ii) medium flows (2 equal segments); iii) high flows. For each segment, there will be a regression equation. Breaking points between segments were determined by finding the intercepts of a linear regression equation and a polynomial equation of second order, both of which were established from gauging data. Figure 3 illustrates our segmentation process and result.

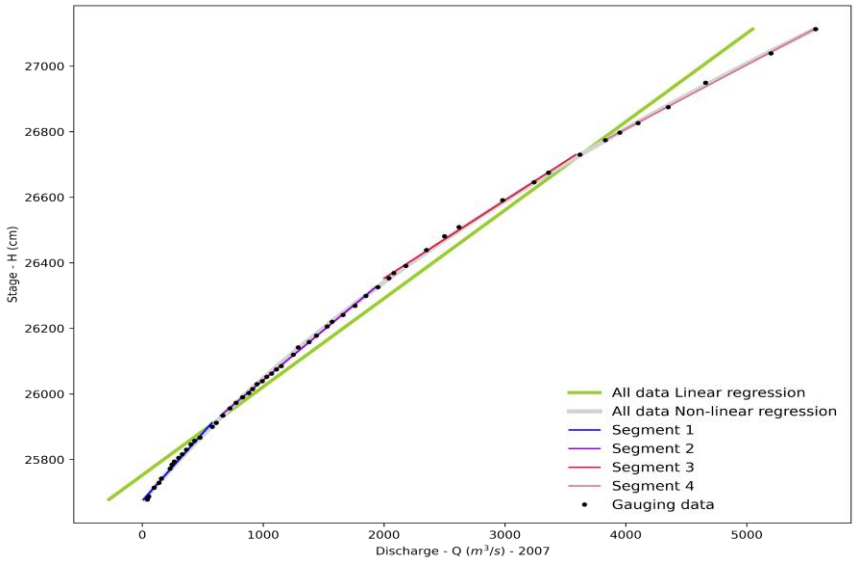

Figure 3. Result of four segments obtained by the proposed automatic segmentation method.

3.3.3. Choosing the best regression equation for discharge interpolation and extrapolation

For discharge interpolation, eight regression equations were established and evaluated by four performance metrics (mentioned in section 3.4), in both piecewise and non–piecewise fashions.

For discharge extrapolation, simple extensions and log extrapolations extend the rating curve using the rating equation for the highest limb of the existing rating curve [18]. The good regression equations of the fourth segment were used for discharge extrapolation. Five and ten percent of gaugings were left out each year. Consequently, each year has an extension ratio, which was calculated using formula (9):

$$ER = \frac{H^{g}_{max} - H^{g}_{cut}}{H^{g}_{max} - H^{re}_{min}} 100\ (\%) \quad (9)$$

where ER (%) is an extension ratio; H_{max}^{g} is the maximum gauging stage data, H_{min}^{re} is the minimum recorded stage data; H_{cut}^{g} is the maximum gauging stage data after high stage data being cut off.

The scatter plots of the ER data on the horizontal axis and the σ or MAE of piecewise ratings in the vertical axis were established and evaluated to arrive at the best model for discharge extrapolation.

3.3.4. Assessing the capability of the combined rating curve in simulating single year discharge

The combined rating curve was created from the regression equations by both Piecewise and Non–piecewise methods, which were computed from gauging data over many years (in this case, 7 years). These combined rating curves interpolate the rating discharge for each year. The discharge simulated from the combined rating curve and the year–specific rating curve were compared to each other as well as to the observed discharge based on performance metrics. These metrics reflected the capacity of the combined rating curve to simulate single–year discharges.

3.4. Performance metrics

To determine the goodness–of–fit of the regression equations this study used four performance metrics, namely KGE, MAE, Pbias and (σ), that are subsequently described as the following.

Kling Gupta efficiency (KGE) [41] can be used to assess the goodness–of–fit between model outputs such as water level, flow data, and climatic data with observed data. KGE values vary in the range from –∞ (not fit at all) to 1 (best fit). KGE is calculated by the following equations:

$$KGE = 1 - \sqrt{(CC-1)^2 + \left(\frac{P_i}{O_i} - 1\right)^2 + \left(\frac{Pm_i}{Om_i} - 1\right)^2};\ CC = \frac{\sum_{i=1}^{n}(O_i - O_m)*(P_i - P_m)}{\sqrt{\sum_{i=1}^{n}(O_i - O_m)^2} * \sqrt{\sum_{i=1}^{n}(P_i - P_m)^2}} \quad (10)$$

The MAE (Mean absolute error) is a metric often used to evaluate how much deviation of a simulated variable from an observed one. The advantage of using this metric is the ease of interpretation since it has the same unit as a variable's unit of interest. But the MAE is not sensitive to outliers. MAE value ranges between 0.0 (best goodness–of–fit) and infinitive (worst goodness–of–fit). The lower the MAE, the better the model fits gauging data [42].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|O_i - P_i| \quad (11)$$

The σ is used as the Vietnamese national standard's metric [36] to evaluate the stability of the Stage – Discharge relationship and the goodness–of–fit of regression models. The model is a good approximation of observed/gauging data when σ < 5%. This metric is formulated as the following:

$$\sigma = \sqrt{\frac{\sum_{1}^{n}\left(\frac{O_i - P_i}{P_i} * 100\%\right)^2}{n}} \quad (12)$$

The percent bias (PBias) can determine simulation bias (negative or positive change) in percentage. The Pbias cannot be applied for single event simulations or can reflect wrong results with a short data. The range value of Pbias is from –100 % to 100%, where the optimal value is 0.0% [42]. This metric can be calculated by using the following equation:

$$Pbias = \frac{\sum_{1}^{n}(O_i - P_i)}{\sum_{1}^{n} O_i} * 100\ (\%) \quad (13)$$

where O_i is observed data; P_i is simulated data; O_m, P_m are the mean of observed and simulated data; n the number of observed data points.

3.5. Building Python script

The mathematical formulations presented from sections 3.1 to 3.4 were codified as a script using Python programming language. The Gradient Descent algorithm was used to find the optimal parameters for each type of regression equation. The mean square error was used as the cost function. The following section presents the results obtained when applying this script to the case study at Polech station, on the Da River.

4. Result and discussion

4.1. Outliers detection and stability of the Q–H relation

4.1.1. Outlier detection

Outlier testing was carried out for all 7 years of gauging data (Figure 4). There is one outlier in the 2005 data set that has been detected using the interquartile range. This outlier reaches outside the upper threshold (U) of water level data and the riverbanks of PoLech station. The remaining data could be utilized for the following steps after all outliers have been removed.

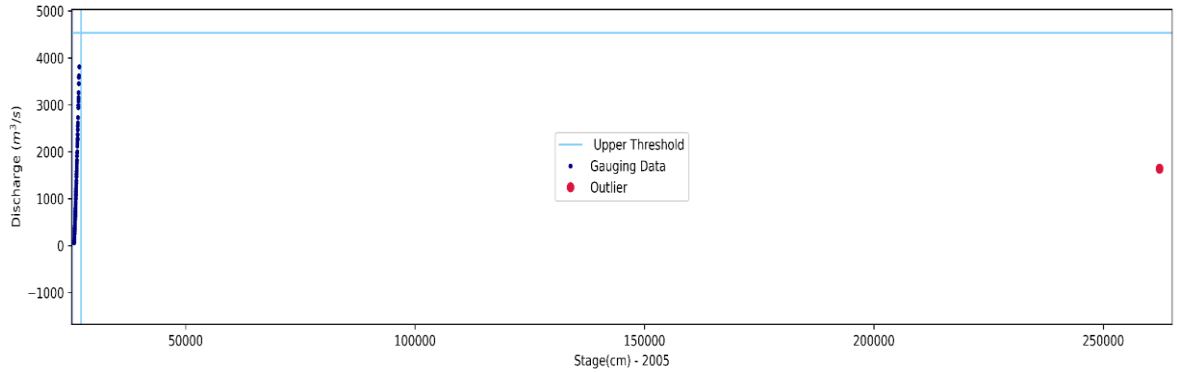

Figure 4. Detecting outlier (red point).

4.1.2. The stability of the Stage – Discharge relationship

Figure 5 depicts the seven rating curves for seven years from 2005 to 2011. There is not much of a distinction among these rating curves. The stage–discharge relationship of PoLech station is relatively steady over time.

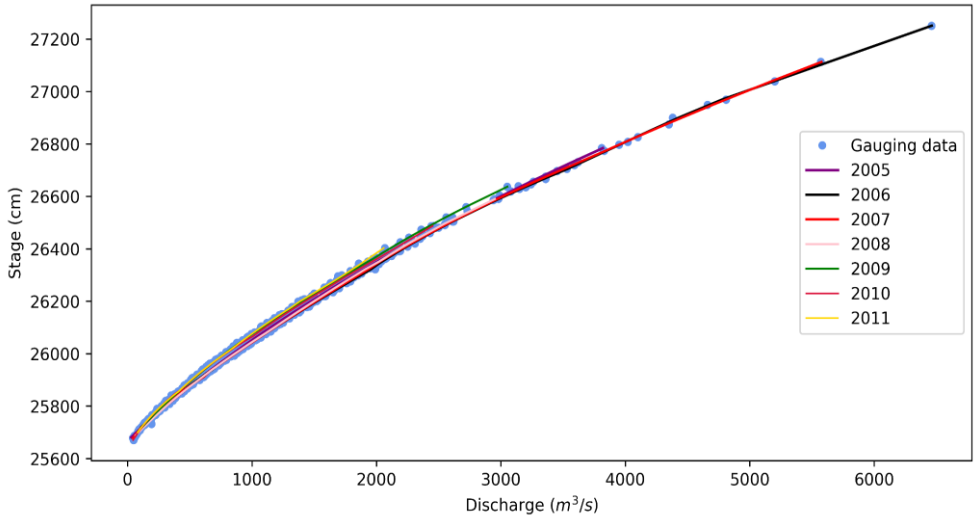

Figure 5. Gaugings (represented by dots) and each–year rating curve (represented by lines) at the PoLech station.

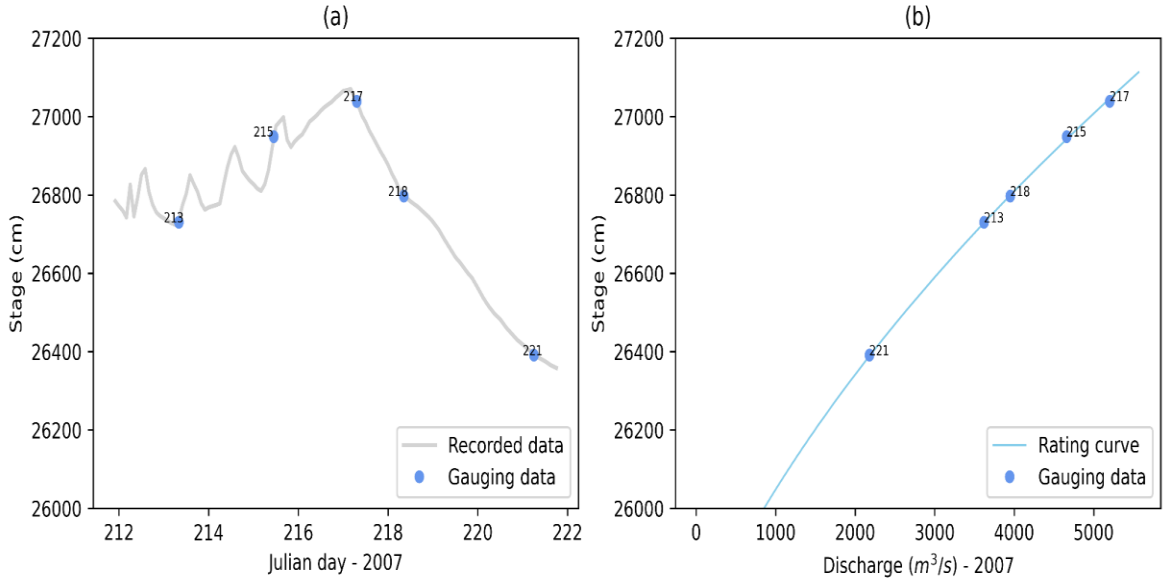

Figure 6. (a) Gauging and recorded flood data from 212 to 221 Julian day in 2007 and (b) the corresponding Stage – Discharge flood rating curve.

Figure 6a plots the gauging data on both rising and falling climbs of a flood hydrograph and all these gauging data fall in a smooth rating curve (Figure 6b). The relationship between stage and discharge is therefore steady and is not affected by flood events. In section 4.2, the performance metrics such as PBIAS, σ give more detail in quantitative analysis for the stability relationship of stage – discharge in PoLech station, which ensures the standard as described in subsection 3.2.

4.2. Discharge interpolation

4.2.1. Interpolation using non–piecewise regression

Table 4 presents four performance metrics resulting from non–piecewise regression equations for discharge interpolation. Some conclusions can be inferred from Table 4, which are: (1) the log–log equation (Eq. 3) is less effective for interpolation than other equations; (2) the non–linear regression equations give better results than linear regression equations; (3) Among the four non–linear equations (Eq. 5 to Eq. 8), it is very hard to rank which one is the best since a particular regression equation is the best in one metric but not in the others.

The rating curves using Eq. 6 and Eq. 7 in the years 2007 and 2010 have very large errors (in term of σ) as compared with those in other years (Table 4). A closer examination of the rating curve using Eq. 6 was carried out by plotting this and the rating curve using Eq. 8 (which has the smallest σ in 2010) against gauging data in Figure 7. The Eq. 6 rating curve is closer to the gauging data for medium and high stage – discharge values in comparison with one using Eq. 8. It is interesting to note that the year 2007 and 2010 have the smallest minimum recorded and gauging stages (Table 1) in which the year 2010 has the smallest recorded stage. For both years, Eq. 8 rating curves closely matched the very low observed data. This result suggested that Eq. 8 should be used for interpolating very low discharge values.

Furthermore, most of the errors of the rating curves (σ) shown in Table 4 are greater than 5%. Therefore, a piecewise regression approach was used, and the results are presented in the next subsection.

Table 4. Performance metrics resulting from non–piecewise regression equations for interpolation.

		Linear				Non – linear			
		Eq. 1	**Eq. 2**	**Eq. 3**	**Eq. 4**	**Eq. 5**	**Eq. 6**	**Eq. 7**	**Eq. 8**
2005	MAE	81.88	81.88	399.88	94.99	22.21	18.68	18.68	44.75
	KGE	0.989	0.989	–0.309	0.878	0.998	0.999	0.999	0.997
	PBias(%)	0	0	–7.440	0.226	–0.034	0	0	0.041
	σ (%)	66.86	66.86	34.04	17.57	5.9	12.45	12.45	8.1
2006	MAE	121.48	121.48	508.54	105.28	29.93	16.60	16.60	53.89
	KGE	0.9796	0.9796	–1.912	0.852	0.997	0.999	0.999	0.997
	PBias(%)	0	0	–12.400	0.246	–0.081	0	0	0.061
	σ (%)	540.13	540.09	30.36	15.56	7.87	2.19	2.19	8.42
2007	MAE	135.22	135.22	604.20	121.50	27.98	16.28	16.28	59.77
	KGE	0.984	0.984	–0.965	0.867	0.997	0.999	0.999	0.997
	PBias(%)	0	0	–12.500	0.297	–0.069	0	0	0.066
	σ (%)	91.9	91.9	39.2	20	10.37	19.23	19.23	11.47
2008	MAE	45.72	45.72	264.94	69.64	9.60	11.47	11.47	25.45
	KGE	0.993	0.993	–0.247	0.882	0.999	0.999	0.999	0.996
	PBias(%)	0	0	–6.530	0.227	–0.007	0	0	0.038
	σ (%)	85	85	29.99	15.79	5.31	8.93	8.93	8.74
2009	MAE	56.37	56.37	221.22	52.55	8.51	10.52	10.52	15.47
	KGE	0.989	0.989	–0.015	0.907	0.999	0.999	0.999	0.998
	PBias(%)	0	0	–4.540	0.108	–0.005	0	0	0.015
	σ (%)	79.36	79.36	22.66	10.79	2.71	4.39	4.39	5.13
2010	MAE	55.83	55.83	255.22	54.82	9.49	14.78	14.78	13.76
	KGE	0.986	0.986	–0.732	0.862	0.999	0.999	0.999	0.998
	PBias(%)	0	0	–8.320	0.181	0.023	0	0	0.011
	σ (%)	183.91	183.91	29.33	14.41	8.3	83.98	83.98	4.07
2011	MAE	33.54	33.54	148.22	38.54	8.20	8.98	8.98	12.98
	KGE	0.993	0.993	0.342	0.916	0.999	0.999	0.999	0.999
	PBias(%)	0	0	–2.270	0.078	0.001	0	0	0.008
	σ (%)	168.93	168.93	17.67	8.23	1.95	3.91	3.91	2.35
Average	MAE	75.72	75.72	343.18	76.76	16.56	13.90	13.90	32.29
	KGE	0.988	0.988	–0.548	0.881	0.998	0.999	0.999	0.998
	PBias(%)	0	0	–7.710	0.195	–0.025	0	0	0.034
	σ (%)	173.72	173.72	29.04	14.62	6.06	19.3	19.3	6.9

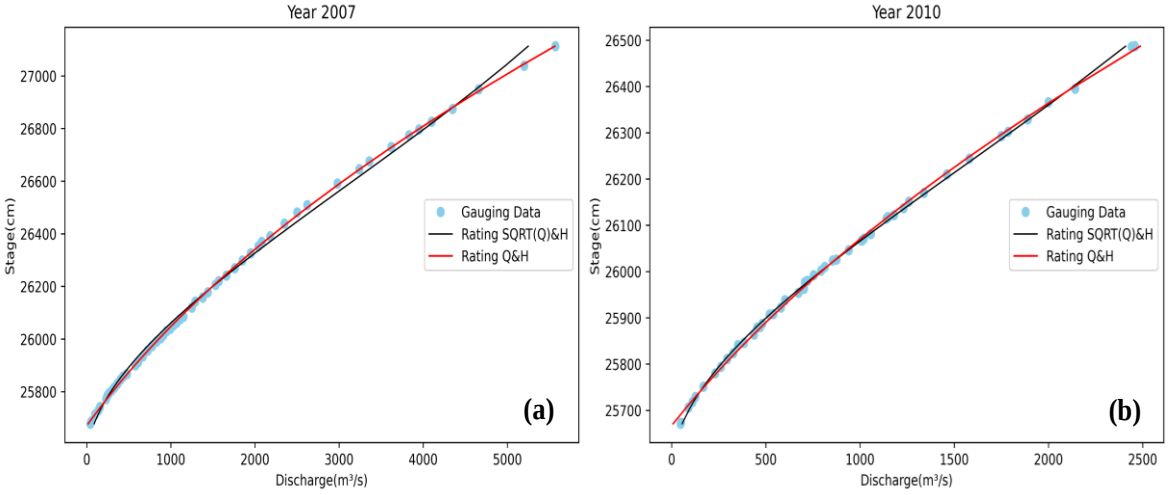

Figure 7. Eq.6 and Eq.8 non–piecewise rating curves in 2007 (a) and 2010 (b).

4.2.2. Interpolation using piecewise regression

The performance metrics shown in Table 5 indicate that piecewise non–linear regression equations give excellent results (lower MAE, KGE close to 1, PBias mostly equal 0, σ less than 5%). Comparing performance metrics in Table 4 and Table 5, it can be concluded that the piecewise method is better than the non–piecewise method, and the non–linear equation is more suitable than linear equation for discharge interpolation.

Table 5. Performance metrics resulting from piecewise regression equations for interpolation.

		Linear				Non – linear		
		Eq. 1	**Eq. 2**	**Eq. 3**	**Eq. 4**	**Eq. 6**	**Eq. 7**	**Eq. 8**
2005	MAE	14.59	14.59	25.21	15.22	12.60	12.60	12.37
	KGE	0.99960	0.99960	0.99823	0.99959	0.99968	0.99968	0.99966
	PBias(%)	0	0	–0.081	0.005	0	0	0.003
	σ(%)	10.97	10.97	7.04	2.48	2.63	2.63	1.69
2006	MAE	15.69	15.69	31.68	14.82	7.98	7.98	8.43
	KGE	0.99962	0.99962	0.99595	0.99947	0.99989	0.99989	0.99989
	PBias(%)	0	0	–0.095	0.006	0	0	0.002
	σ(%)	3.49	3.49	6.38	2.74	1.13	1.13	1.16
2007	MAE	14.64	14.64	33.24	13.29	7.59	7.59	8.29
	KGE	0.99979	0.99979	0.99713	0.99929	0.99994	0.99994	0.99995
	PBias(%)	0	0	–0.166	0.005	0	0	0.002
	σ(%)	42.41	42.41	10.32	4.20	1.51	1.51	1.99
2008	MAE	10.18	10.18	17.29	9.85	7.68	7.68	7.90
	KGE	0.99969	0.99969	0.99708	0.99921	0.99979	0.99979	0.99985
	PBias(%)	0	0	–0.046	0.013	0	0	0.009
	σ(%)	4.97	4.97	8.87	6.17	4.89	4.89	4.96
2009	MAE	10.10	10.10	14.14	8.01	7.23	7.23	7.19
	KGE	0.99967	0.99967	0.99837	0.99976	0.99979	0.99979	0.99983
	PBias(%)	0	0	–0.031	0.004	0	0	0.003
	σ(%)	2.48	2.48	4.97	2.83	1.98	1.98	2.04
2010	MAE	9.83	9.83	12.57	7.61	7.09	7.09	6.87
	KGE	0.99960	0.99960	0.99863	0.99982	0.99977	0.99977	0.99979
	PBias(%)	0	0	–0.073	0.004	0	0	0.003
	σ(%)	10.90	10.90	6.12	1.88	2.00	1.99	1.50
2011	MAE	9.21	9.21	10.05	7.60	7.19	7.19	7.26
	KGE	0.99931	0.99931	0.99805	0.99969	0.99950	0.99950	0.99950
	PBias(%)	0	0	–0.022	0.005	0	0	0.004
	σ(%)	3.59	3.59	3.29	1.68	1.55	1.55	1.54
Average	MAE	12.03	12.03	20.60	10.91	8.19	8.19	8.33
	KGE	0.99961	0.99961	0.99763	0.99955	0.99977	0.99977	0.99978
	PBias(%)	0	0	–0.074	0.006	0	0	0.004
	σ(%)	11.26	11.26	6.71	3.14	2.24	2.24	2.13

To further illustrate the effectiveness of the piecewise regression, both piecewise rating curves using Eq. 6 and Eq. 8 in 2007 and 2010 are shown in Figure 8. As compared with the corresponding rating curves using non–piecewise regression (Figure 7), the automatic segmentation method for piecewise regression proposed in this study substantially improved the goodness–of–fits of the two regression equations.

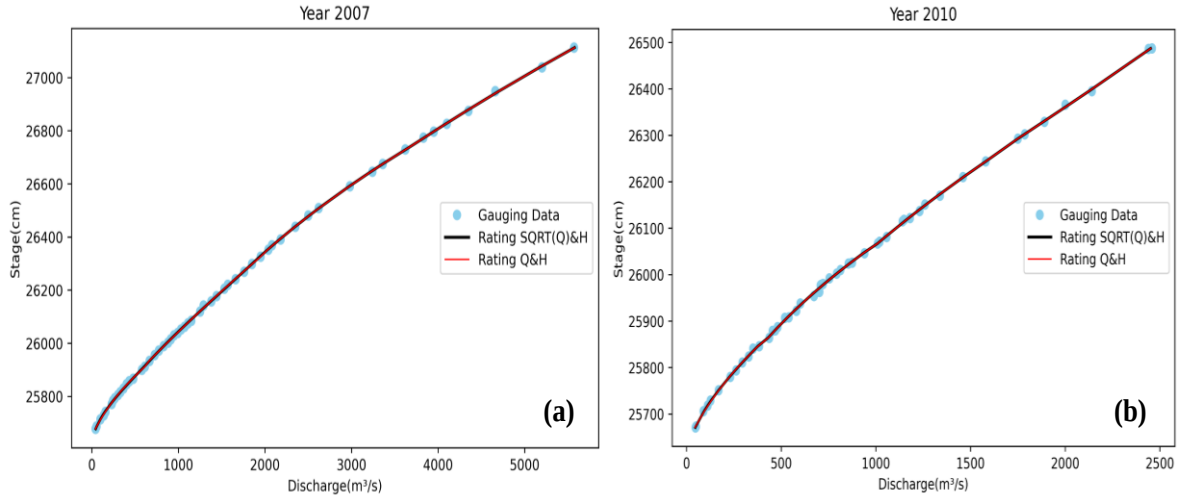

Figure 8. Eq. 6 and Eq. 8 piecewise rating curves in 2007 and 2010.

Regarding selection of the best equation for interpolation, all the three non–linear equations give very similar and satisfactory results. Therefore, either of them can be selected for the purpose of interpolation, given that automatic piecewise regression has been applied.

4.3. Applying the combined rating curve to interpolate discharge for each year

For discharge interpolation, the non–linear regression gives better results in both piecewise and non–piecewise methods. Therefore, to assess the ability of the combined rating curve (being built using all gauging data from 2005 to 2011) for individual year discharge interpolation, the non–linear regression equations (Eqs. 6, 7, 8) were used. Four performance metrics resulting from regressing these equations are presented in Table 6. Inferring from Table 6 is that piecewise regression improved the σ (approaching 5 %), but not for other metrics. Applying the combined rating curve to interpolate each year's discharge requires more research.

Table 6. Seven–year averaged performance metrics computed by the Combined rating curve (Non – linear regression).

		MAE	KGE	PBias(%)	σ (%)
Non piecewise	Eq. 6	33.26	0.95589	–0.583	17.48
	Eq. 7	33.26	0.95589	–0.583	17.48
	Eq. 8	47.31	0.90867	–0.364	9.75
Piecewise	Eq. 6	37.04	0.95656	–0.825	5.95
	Eq. 7	37.04	0.95656	–0.825	5.95
	Eq. 8	37.45	0.95140	–1.014	5.28

4.4. Extrapolation of flow at high water level

Table 7 presents the analysis results using both linear and non–linear regressions for flow extrapolation. In overall, the linear regression using Eq. 4 outperformed the other equations, and the Eq. 3 performed the worst. Except for the log–log relation (Eq. 3), the linear regression did a better job than the non–linear regression.

Table 7. Performance metrics and the extension ratios in 7 years.

Cut–off percentage	Year	ER (%)	Linear						Non–linear			
			Eq. 1		Eq. 3		Eq. 4		Eq. 6		Eq. 8	
			MAE	σ (%)	MAE	σ (%)	MAE	σ (%)	MAE	σ (%)	MAE	σ (%)
5%	2010	11.27	7.19	0.38	130.86	5.08	61.62	2.47	46.29	1.88	41.94	1.71
	2005	11.6	45.19	1.42	67.58	2.36	28.58	1.02	29.33	0.97	29.61	0.95
	2007	16.59	60.06	1.19	226.53	4.32	81.83	1.63	122.53	2.37	131.51	2.54
	2009	17.6	52.01	2.06	77.26	2.9	5.67	0.22	147.49	6.16	176.80	7.6
	2011	18.21	32.74	1.74	159.95	8.1	87.08	4.53	125.95	6.45	131.37	6.73
	2008	18.85	62.85	2.52	76.74	2.83	3.48	0.13	96.34	3.9	114.47	4.72
	2006	28.94	160.82	3.69	558.11	9.73	150.74	2.84	64.35	1.56	83.91	1.76
10%	2005	16.41	76.52	2.5	49.98	1.96	32.46	1.03	29.46	1.28	37.14	1.57
	2007	22.02	70.67	1.75	200.28	4.56	51.66	1.28	26.02	0.69	22.00	0.57
	2010	22.67	30.57	1.57	137.24	6.55	41.63	2.23	31.42	1.66	29.03	1.51
	2008	24.94	31.43	1.5	116.37	4.74	46.90	1.88	130.46	5.25	151.39	6.13
	2009	27.15	27.82	1.39	155.02	6.08	55.10	2.16	41.54	1.63	44.32	1.74
	2011	29.02	47.81	2.99	100.48	6.45	33.31	2.04	33.81	2.25	43.08	2.71
	2006	40.24	242.92	6.48	471.23	10.19	51.89	1.26	191.05	4.5	238.44	5.59
Average			**67.76**	**2.23**	**180.55**	**5.42**	**52.28**	**1.77**	**79.72**	**2.9**	**91.07**	**3.27**

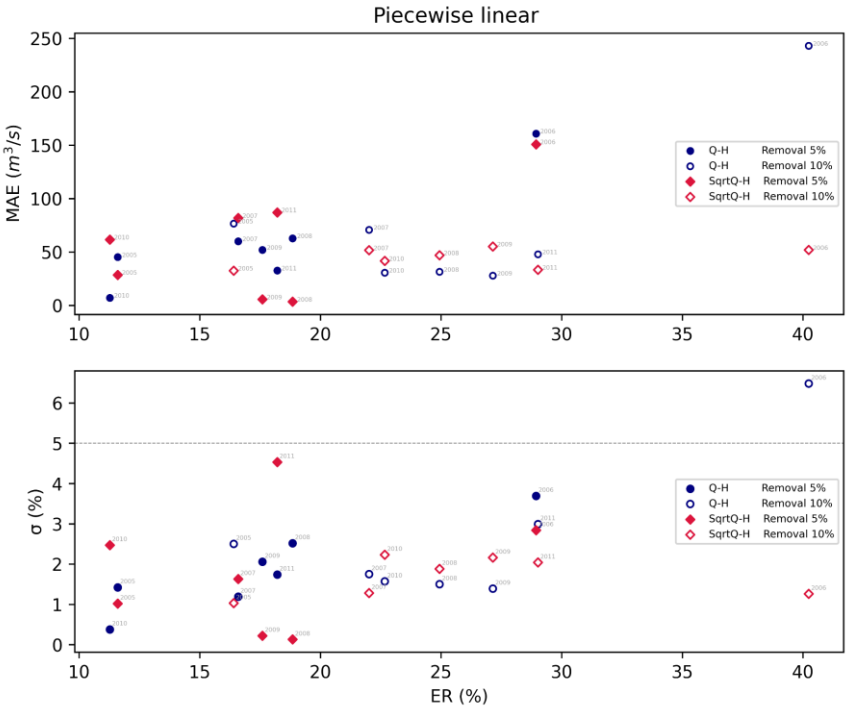

Figure 9. Scatterplots of σ against MAE (the upper) and ER (the lower) computed by two linear regression equations (Eq. 1 and Eq. 4).

In Fig. 9, the scatterplots of σ against MAE (the upper) and ER (the lower) computed by two linear regression equations for seven years (each year has two paired values corresponding to 5% and 10 % cut–off rate). In 2006, although the highest gauging discharge was 6460 m^3/s, the Eq. 4 still gives excellent results when the extension ratio is of more than 40%. For all the other pairs, both linear equations give very satisfactory results (considering both σ and MAE) even when ER goes up to 30%. The smallest discharge value needed to be extrapolated in all of seven years is 1420 m^3/s (in 2011), and all values of MAE of pairs are smaller than 100 m^3/s. Only one exception is 2006 where discharge of 6460 m^3/s needed to be found. According to [18], the error of minus/plus 10 % for high discharge extrapolation is exceptionally ideal, and this is the case in this study.

5. Concluding remarks and Outlooks

In this research, eight regression equations (Equations in Table 2) were evaluated and compared to determine the most suitable rating curve equation for interpolating discharge and extrapolating high–flow rates with the use of water level (stage) measured at the PoLech station in the Da river. To aid in the process, a straightforward segmentation approach was employed. The outcome of the comprehensive investigations revealed that: (1) For a single–year dataset, the second–order polynomial regression equations (in which stage is the independent variable and either Q or $Q^{1/2}$ is the dependent variable) demonstrated superior performance in capturing the non–linearity of the data as compared to other approaches for interpolating the discharge. (2) For extrapolation, the linear regression equation was suggested as the most appropriate. (3) For all–year data (i.e., 7 years), the combined rating curve produced discharge estimations that were less accurate than those from a single–year rating curve. The potential of being able to produce ongoing discharge estimations at minimal expense and with relatively straightforward calibration techniques is far-reaching. From an operational standpoint, this approach may stimulate researchers, aquatic eco–system stewards, water quality monitors, or assessors of upstream withdrawals to begin gauging river discharge on a more frequent basis.

The work is not without limitations. First, further efforts should be invested in the evaluation of utilizing a combined rating curve for each year's discharge interpolation. Observations acquired from the current case study demonstrate that the shifting of rating curves over time (years) should be considered even though for a specific year, the hydraulic regime is regarded as constant and stable. Second, although many regression experiments were conducted in this study, other equations, such as spline and Chebyshev polynomial, should be tested using this case study data in a piecewise manner to determine if improved outcomes can be attained. Third, in this study, our goal was to find the most suitable equation based on the best–fitted rating curve, so the uncertainty was not considered. It is important to note that numerous factors influence rating curve uncertainty estimation regardless of the modeling approach, such as the non–linearity relationship between the water stage and discharge or the alteration of the riverbed. Investigating the uncertainty of these factors would be our next step. Lastly, although seven year–station datasets were utilized in this study, more datasets obtained from other stations will help to increase the generality of our approach.

Author contribution statement: Designed the study conception: G.N.T.; collected data: G.N.T., T.N.T.; developed the theoretical research: G.N.T., H.N.D.; processed the data and performed the calculations: M.D.T.D., H.D.B., Q.H.D.; analyzed the data: G.N.T., H.N.D., M.D.T.D., H.D.B., Q.H.D.; contributed largely to revising the final manuscript: G.N.T., V.T.N.

Acknowledgements: This study is supported by project No. ĐTĐL.CN–06/23 of the 562–programme funded by Vietnam Ministry of Science and Technology.

Conflicts of interest: All authors have no conflicts of interest to declare in this research.

References

1. ISO 7066-2:1988. Assessment of uncertainty in the calibration and use of flow measurement devices – Part 2: Non–linear calibration relationships. In: International Standard 7066–2, International Organization for Standardization, Geneva, 1988.
2. Karimi, P.; Bastiaanssen, W.G.M. Spatial evapotranspiration, rainfall and land use data in water accounting–Part 1: Review of the accuracy of the remote sensing data. *Hydrol. Earth Syst. Sci*. **2015**, *19(1)*, 507–532.
3. Stewart, B. Measuring what we manage–the importance of hydrological data to water resources management. Proceedings of the International Association of Hydrological Sciences **2015**, *366*, 80–85.
4. Xu, C.y. Issues influencing accuracy of hydrological modeling in a changing environment. *Water Sci. Eng*. **2021**, *14(2)*, 167–170.
5. Zakwan, M.; Muzzammil, M.; Alam, J. Application of data driven techniques in discharge rating curve–an overview. *Aquademia* **2017**, *1(1)*, 02.
6. Leopold, L.B. River channel change with time: an example: address as retiring president of the Geological Society of America, Minneapolis, Minnesota, November 1972. *Geol. Soc. Am. Bull*. **1973**, *84(6)*, 1845–1860.
7. Booth DB. Stream-channel incision following drainage-basin urbanization 1. *JAWRA J. Am. Water Resour. Assoc.* **1990**, *26(3)*, 407–417.
8. Jennings, D.B.; Taylor Jarnagin, S. Changes in anthropogenic impervious surfaces, precipitation and daily streamflow discharge: a historical perspective in a mid–Atlantic subwatershed. *Landsc Ecol*. **2002**, *17*, 471–489.
9. Ajami, N.K.; Hornberger, G.M.; Sunding, D.L. Sustainable water resource management under hydrological uncertainty. *Water Resour Res*. **2008**, *44(11)*, W11406.

10. Leopold, L.B. Hydrology for Urban Land Planning: A Guidebook on the Hydrologic Effects of Urban Land Use. US Geolgoical Survey, 1968, 554, pp. 18.
11. Wharton, G.; Tomlinson, J.J. Flood discharge estimation from river channel dimensions: results of applications in Java, Burundi, Ghana and Tanzania. *Hydrol. Sci. J.* **1999**, *44(1)*, 97–111.
12. Wara, C.; Thomas, M.; Mwakurya, S.; Katuva, J. Development of River Rating Curves for Simple to Complex Hydraulic Structure Based on Calibrated HEC–RAS Hydraulic Model, in Kwale, Coastal Kenya. *J. Water Resour. Prot.* **2019**, *11(04)*, 468.
13. Sivapragasam, C.; Muttil, N. Discharge rating curve extension–a new approach. *Water Resour. Manage.* **2005**, *19*, 505–520.
14. Fenton, J.D.; Keller, R.J. The calculation of streamflow from measurements of stage. Technical Report 01/6, 2001, pp. 77.
15. Venetis, C. A note on the estimation of the parameters in logarithmic stage–discharge relationships with estimates of their error. *Hydrol. Sci. J.* **1970**, *15(2)*, 105–111.
16. McMillan, H.; Krueger, T.; Freer, J. Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality. *Hydrol. Process.* **2012**, *26(26)*, 4078–4111.
17. Petersen–Øverleir, A. Modelling stage—discharge relationships affected by hysteresis using the Jones formula and nonlinear regression. *Hydrol. Sci. J.* **2006**, *51(3)*, 365–388.
18. Ramsbottom, D.M.; Whitlow, C.D. Extension of Rating Curves at Gauging Stations Best Practice Guidance Manual. Environment Agency, Rio House, Waterside Drive, Aztec West, Almondsbury, Bristol, BS32 4UD, 2003, pp. 247.
19. Holmes Jr, R.R. River rating complexity. Constantinescu, editor, River Flow. Proceeding of the International Conference on Fluvial Hydraulic (River Flow 2016), St. Louis, Missouri, July 11-14, 2016: CRC Press, 2016, 679–686.
20. Léonard, J.; Mietton, M.; Najib, H.; Gourbesville, P. Rating curve modelling with Manning's equation to manage instability and improve extrapolation. *Hydrol. Sci. J.* **2000**, *45(5)*, 739–750.
21. Domeneghetti, A.; Castellarin, A.; Brath, A. Assessing rating–curve uncertainty and its effects on hydraulic model calibration. *Hydrol. Earth Syst. Sci.* **2012**, *16(4)*, 1191–1202.
22. Herschy, R.W. Streamflow Measurement. *CRC Press*. 1995, pp. 536. https://doi.org/10.1201/9781482265880.
23. Petersen–Øverleir, A.; Reitan, T. Objective segmentation in compound rating curves. *J. Hydrol. (Amst).* **2005**, *311(1–4)*, 188–201.
24. Zarzer, E.A. Some considerations concerning the optimal calculation of stage–discharge functions. *Zeitschrift für Oper. Res.* **1987**, *31(6)*, B193–B212.
25. Petersen–Øverleir, A. Accounting for heteroscedasticity in rating curve estimates. *J Hydrol. (Amst).* **2004**, *292(1–4)*, 173–181.
26. Moyeed, R.A.; Clarke, R.T. The use of Bayesian methods for fitting rating curves, with case studies. *Adv. Water Resour.* **2005**, *28(8)*, 807–818.
27. Aggarwal, S.K.; Goel, A.; Singh, V.P. Stage and discharge forecasting by SVM and ANN techniques. *Water Resour. Manage.* **2012**, *26*, 3705–3724.
28. Londhe, S.; Panse–Aglave, G. Modelling stage–discharge relationship using data–driven techniques. *ISH J. Hydraul. Eng.* **2015**, *21(2)*, 207–215.
29. Fenton, J.D. On the generation of stream rating curves. *J. Hydrol. (Amst).* **2018**, *564*, 748–757.
30. Herschy, R.W. Streamflow Measurement. *CRC Press* 2008, pp. 536.

31. Morgenschweis, G. Hydrometrie: Theorie Und Praxis Der Durchflussmessung in Offenen Gerinnen. Springer Vieweg Berlin, Heidelberg, 2010, XVI, pp. 637.
32. Mirza, M.M.Q. The choice of stage–discharge relationship for the Ganges and Brahmaputra rivers in Bangladesh. *Hydrol. Res*. **2003**, *34(4)*, 321–342.
33. McMahon, T.A.; Peel, M.C. Uncertainty in stage–discharge rating curves: application to Australian Hydrologic Reference Stations data. *Hydrol. Sci. J*. **2019**, *64(3)*, 255–275.
34. Khai, N.H.; Cau, L.V. Results of research on the stable rating curve in the tidal unaffected area using the function Spline 3. *VN J. Hydrometeorol*. Published online 1996.
35. Dao, N.V.; Hai, Hai, L.Q.; Ky, N.D.; Phong, P.H.; Dat, D.V.; Phung, N.V.; Chien, L.Q. Study on building correlation of water level and discharge at Ha Bang hydrological station in the period 2013–2020. *VN J. Hydrometeorol*. **2021**, *723*, 38–47.
36. Vietnamese national standard TCVN 12636–15:2021. Hydro–Meteorological Observations – Part 15: Editing of water flow documents discharge in river on non – tidal affected zones. 2021.
37. Anh, T.V.; Hien, N.T.; Khanh, D.Q. Estimating the foreign flow from China to Vietnam supporting water resources planning and management in Da river basin. *VN J. Hydrometeorol*. **2017**, *678*, 54–62.
38. Linh, B.H.; Phuong, T.A. Assessment of the impact of reservoirs on flow variations on the Da River. *VN J. Hydrometeorol*. **2021**, *731*, 97–107.
39. Ministry of natural resources and environment of the Socialist Republic of Vietnam. National Water Resources Report Period 2016–2021, 2022.
40. Barbato, G.; Barini, E.M.; Genta, G.; Levi, R. Features and performance of some outlier detection methods. *J. Appl. Stat*. **2011**, *38(10)*, 2133–2149. doi:10.1080/02664763.2010.545119.
41. Gupta, H.V.; Kling, H.; Yilmaz, K.K.; Martinez, G.F. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol. (Amst)*. **2009**, *377(1–2)*, 80–91.
42. Moriasi, D.N.; Gitau, M.W.; Pai, N.; Daggupati, P. Hydrologic and water quality models: Performance measures and evaluation criteria. *Trans ASABE*. **2015**, *58(6)*, 1763–1785.