MCIT Modeling, control & information technologies

A Design Water Discharge Maxima Forecasting Method Based on Observation Data Using Plotting Position Formulas

https://doi.org/10.31713/MCIT.2023.061

Dmytro Stefanyshyn
Department of Natural Resources
Institute of Telecommunications and Global Information Space of the NAS of Ukraine
Kyiv, Ukraine
d.v.stefanyshyn@gmail.com

Abstract—**Plotting position formulas provide a non-parametric means to estimate the observed hydrological data probability distribution. In particular, using a plotting position formula, a plot of the estimated values from a theoretical parametric probability distribution can be compared with the observation data. It allows an examination of the adequacy of the fit provided by parametric probability distributions. However, results of calculating empirical annual probabilities of exceedance observed maxima water discharges show an increase in the divergence between the estimates obtained using the different plotting position formulas in case of more extreme events. Thereby, the choice of a relevant plotting position formula becomes a challenge. Different plotting position formulas may be admissible options. This article shows that the divergence between the plotting position estimates can be extrapolated to predict design maxima water discharges of low exceedance probabilities.**

Keywords—extrapolation; forecasting; observation data; plotting position formulas; uncertainty.

I. INTRODUCTION

There are two basic approaches to forecasting in modern hydrology and water management. These are a probabilistic approach and genetic (deterministic) means. The genetic methods are more adequate and scientifically correct but complicated to realize in practice because, in the majority, river runoff reasons and processes are complex, multifactorial, interrelated, and stochastic. Therefore, the probabilistic methods based on hydrological observation data statistic analysis have been more popular in studying and predicting diverse design hydrological characteristics. In particular, it is in the case of forecasting maxima water discharges relating to riverine floods [1-3].

The probabilistic approach is based on frequency analyses of observation data [4]. Usually, the time series of observed maxima water discharges are considered and analyzed in the frame of the stationary hypothesis. To forecast design discharges of low annual probability of exceedance, parametrical probability distributions are used as predictive models [2]. Returning to the problem of flood risk management, it should be reminded that Directive 2007/60/EC (the EU Flood Risk Directive) [5] defines flood risk quantitatively as “the combination of the probability of a flood event and the potential adverse consequences for human health, the environment, cultural heritage and economic activity associated with a flood event”. Thus, in any case, the quantitative flood risk assessment will require the calculation of the probabilities (frequencies) of observed maximal (peak) flood discharges.

Hydrological maxima are specific extreme events. In theory, they are not limited to the upper limit. Usually, time series of observed pick discharges hold an essential positive asymmetry (skewness); sometimes – strong outliers [4]. Hydrologists are aware that the true probability distributions of maxima discharges of rivers are not being identified. In general, there is no theoretical or other proper justification for choosing an appropriate parametric probability distribution to predict peak discharges of floods using observed data [6]. Practice shows different distributions can well fit the observed time series of annual maximum discharges. However, they can forecast various values of peak discharge Q (m^3/s) of a chosen annual probability of exceedance P (1/year). Vice versa, depending on different distributions, the water discharge can have different values of probabilities of exceedance [3, 6-9].

II. THE PROBLEM FORMULATION AND THE OBJECTIVE OF THE PAPER

There are a lot of standardized probability distribution functions [10] that can be possible options (Table I) to choose from, and any of them might be considered a permissible hypothesis [1, 3, 4, 6-9].

TABLE I. STANDARDIZED PROBABILITY DISTRIBUTION FUNCTIONS IN FREQUENCY ANALYSIS OF MAXIMUM DISCHARGES [10]

Probability Distribution Functions	Country
Pearson type III distribution (P3)	China, Switzerland
Logarithmic Pearson type III distribution (LP3)	USA, Canada, India
Extreme value type I, type III distributions (EV1, EV3), Generalized extreme value distribution (GEV)	Germany, Great Britain, France
Two, Three parameters log-normal distribution (LN2, LN3)	Japan
Extreme value type I distribution (EV1)	Sweden, Norway
Kritskyi-Menkel three-parameter distribution (KM3)	Ukraine, former USSR countries

To test them and choose the best option, a plot of the estimated values from a theoretical parametric probability distribution is compared with the observation data. However, the choice of a relevant plotting position formula may also be a challenge when fitting parametric probability distributions [11]. More than seventeen different plotting position formulas have been proposed by hydrologists and statisticians over the years [12]. Some of the plotting position formulas, most frequently appearing in the hydrological literature, are shown in Table II. All these formulas provide a non-parametric means to estimate the observed data probability distribution. However, there is no worthwhile criterion for comparing plotting position formulas to choose the more appropriate one relating to a real case study.

TABLE II. TYPICAL PLOTTING POSITION FORMULAS [6, 11-13]

Author (year)	Formula to calculate P_m (1/year)[a]
Hazen (1914)	$(m – 0.5)/n$
Gringorten (1963)	$(m – 0.44)/(n + 0.12)$
Nguyen (1989)	$(m – 0.42)/(n + 0.3C_S + 0.05)$
Cunnane (1978)	$(m – 0.4)/(n + 0.2)$
Blom (1954)	$(m – 3/8)/(n + 1/4)$
Hosking (1990)	$(m – 0.35)/n$
Tukey (1962)	$(m – 1/3)/(n + 1/3)$
Goel (1993)	$(m – 0.02C_S – 0.32)/(n – 0.04C_S + 0.36)$
Beard (1945)	$(m – 0.3175)/(n +0.365)$
Kim (2012)	$(m – 0.32)/(n + 0.0149C_S^2 – 0.1364C_S + + 0.3225)$
Lebedev (1952), Chegodaev (1965)	$(m – 0.3)/(n + 0.4)$
Adamowski (1985)	$(m – 0.25)/(n + 0.5)$
Weibull (1939)	$m/(n + 1)$

a. P_m is the empirical probability of exceedance of the m-th order observed value, m is the rank of the value, where the highest one being "1", and n is the number of observed statistics

Perhaps, in time, we will know which of the recommended parametric probability distributions or plotting position formulas were better in a contest of forecasted values of future events. However, it may be checked only after those events happen. In any case, while forecasting, we should regard both the natural (stochastic) uncertainty of observation data and the epistemic (non-stochastic or subjective) uncertainty relating to applying different models [3, 6, 8, 9].

This paper aims to present for discussion a new numerical-analytical method of forecasting design maxima discharges using observation data. It is based on extrapolation of the divergence between the empirical probability estimates that can be obtained by different plotting position formulas. The method is presented in the case of forecasting the maxima discharges of 0.5% and 1% annual probability of exceedance for the Uzh River, the Transcarpathia region, Ukraine, using the hydrological station (HS) "Uzhhorod" observation data from 1947 to 1999.

III. MATERIALS AND METHODS

The study employs a time series of maximum discharges of the Uzh River observed at the HS "Uzhhorod" from 1947 to 1999 [6]. The data sample length is 53 years. The maximum discharge within the data sample is 1680 m^3/s (1957); the minimum – of 146 m^3/s (in 1961). The mean peak discharge is 689 m^3/s; the sample standard deviation – of 364 m^3/s. The coefficient of variation C_V of the time series is 0.53, the skewness C_S is 0.52, and the C_S/C_V is 0.99.

In the study, the scientific methods of theoretical and empirical research, analysis and synthesis, expert evaluation and comparison, formalization and modeling were used, including (1) extrapolation methods [14]; (2) fundamental and practical methods of mathematical statistics [11-13, 15]; (3) specific statistical methods in hydrology [1-4, 6-10]; (4) utility theory methods [16, 17] and decision making methods under risk and uncertainty [8, 9, 18].

To test the proposed forecasting method based on observation data using plotting position formulas, thirteen such formulas were used (See Table II). They were considered in terms of possible expert judgments (suggestions) for assessing the annual empirical probabilities of exceedance of observed maxima discharges. As possible theoretical alternatives for forecasting design maxima discharges of the Uzh River at the HS "Uzhhorod" considered were five parametric distributions: 1) the Kritskyi-Menkel three-parameter distribution (KM3) (C_V = 0.53, C_S = C_V); 2) Pearson's type III distribution (P3) (C_S = 0.52); 3) the Extreme value type I distribution (Gumbell's type I distribution, EV1); 4) the Logarithmic Pearson type III distribution (LP3) (CS = –0.44); and 5) the Two parameters logarithmic-normal distribution (LN2).

IV. THE PRESENTATION OF THE PROPOSED FORECASTING METHOD

A. Preliminary modeling and making assumptions

The pre-modeling included calculating empirical annual probabilities of exceedance P_m observed maxima discharge employing various plotting position formulas depending on the rank $m = 1,...,n$ of the observed values, where the highest one has the rank $m = 1$, and $n = 53$ is the number of observed data. The probabilities P_m were presented as percentages.

Four model cases of forming a data sample were considered. The first model data sample included observed maxima water discharges from 1947 to 1976 (30 years); the second data sample was from 1947 to 1984 (38 years); the third data sample was from 1947 to 1992 (46 years); the fourth (control sample to test the method) was from 1947 to 1992 (53 years).

It was revealed that different plotting position formulas provide similar results for high probable events with short return periods $T_{r,m}$ ($T_{r,m} = 1/P_m$, or $T_{r,m} = 100/P_m$ if P_m is presented as percentages). These events have return periods $T_{r,m}$ of 5 years or less; the annual probabilities of exceedance P_m are 20% and more. However, while enlarging the modeling horizon towards low probable (more extreme) events, the disagreement (divergence) between results tends to increase. The same conclusion applies to the chosen alternative parametric probability distributions.

As a metric of disagreement between probabilities obtained using different plotting position formulas, the divergence indicator d_m was proposed [6]:

$$d_m = P_{m,i} / P_{m,j},\text{ or } d_m = T_{r,m,j} / T_{r,m,i},\ i \neq j, \qquad (1)$$

where $P_{m,i}$, $P_{m,j}$ are the empirical annual probabilities of exceedance of the observed maxima discharges, and $T_{r,m,i}$, $T_{r,m,j}$ are their return periods calculated using the i-th and j-th counterparty plotting position formulas, which provide $P_{m,i} > P_{m,j}$, $T_{r,m,j} > T_{r,m,i}$, and $d_m > 1$ (e. g., Weibull's and Hazen's formulas giving marginal, i.e. maximum and minimum, plotting position probability values), correspondingly; m is a rank of a maxima discharge where the highest one has the rank $m = 1$.

As a result, regression dependencies relating to the indicator d_m can be defined: (a) between the return periods $T_{r,m,i}$, $T_{r,m,j}$ calculated using the i-th and j-th plotting position formulas and the divergence indicator d_m; (b) between the observed maxima water discharges Q_m and the indicator d_m. The regressions $d_{m,i} = f(T_{r,m,i})$ and $d_{m,j} = f(T_{r,m,j})$ indicate that further enlarging of the return period of the observed maxima discharge may correspond to an increase in the divergence in plotting position estimates the different formulas provide. This disagreement depends on the plotting position formulas chosen to be compared. The regression $d_m = f(Q_m)$ indicates that further enlarging of the observed maxima discharge may also correspond to an increase in the divergence in plotting position estimates the different formulas provide. By estimating the indicator d_m and building these regressions, we can make predictions by applying extrapolation. In the first step, the prediction is implemented using the direct dependencies between the disagreement indicator values and the design discharge return period values. In the second step, it is used the dependence between the discharges and the divergence indicator values. Predicting design discharges is made using an iterative calculation method.

B. *Using the Fishburn rule to overcome epistemic uncertainty of plotting position formulas*

Results obtained using different plotting position formulas may be considered expert estimates [6]. Under decision-making, these expert judgments may acquire different importance [16-18]. For example, in flood management strategies, the plotting position estimates obtained according to Weibull contribute to choosing more cautious decision options. However, more cautious options can be associated with additional capital costs. Using Hazen's plotting position formula contributes to choosing options with lower capital costs. In turn, it may inflict an increase in flood losses in future.

When making decisions, different plotting position formulas can be considered indicators of the predisposition to more cautious or less expensive decision options. In other words, various plotting position estimates obtained using different plotting position formulas can acquire their weight level in a system of indicators' importance under the decision-making process [6].

An optimal distribution of the weights of the indicators from the point of view of informational entropy is referred to as Fishburn's rule. The Fishburn rule considers that the level of indicators' importance is determined only by arranged in descending order of importance [16, 17]. According to this rule, the "weight" w_i for the i-th plotting position estimate $P_{m,i}$ obtained using the i-th formula can be calculated as:

$$w_i = 2\,(k - i + 1)\,/\,(k + 1)k, \qquad (2)$$

where i is the rank of the i-th plotting position estimate obtained using the i-th formula taking into account the level of the formula importance; the highest estimate gets the rank $i = 1$ when there is a predisposition to more cautious options, and, vice-versa, when there is a predisposition to options with lower capital costs, the smallest one has the rank $i = 1$; k is the total number of the ranked-set plotting position estimates (formulas).

As a result, depending on the selected significance option of the different plotting position formulas, the rank-weighted estimate of the annual plotting position probability $P_{m,w}$:

$$P_{m,w} = \sum P_{m,i} \cdot w_i\,,\ i = 1, \ldots, k, \qquad (3)$$

where m is the rank of an observed maxima water discharge Q_m (m^3/s).

Depending on the selected significance option of the different plotting position formulas, using the Fishburn rule enables getting two possible rank-weighted estimates of the annual plotting position probability $P_{m,w}$: the rank-weighted upper bound estimate (sup) $P_{m,w,sup}$, the rank-weighted lower bound estimate (inf) $P_{m,w,inf}$. The rank-weighted upper bound estimate $P_{m,w,sup}$ will correspond to the predisposition to more cautious decision options. The rank-weighted lower bound estimate $P_{m,w,inf}$ will correspond to the predisposition to less expensive decision options.

Table III shows the results of the forecasting using the proposed method. These are values of the design maxima discharges of 1% and 0.5% annual probabilities of exceedance for the Uzh River, the HS "Uzhhorod". Forecasting was carried out for four model cases of forming data samples.

TABLE III. RESULTS OF FORECASTING THE DESIGN MAXIMA DISCHARGES (P = 1 AND 0.5 1/YEAR, %), THE UZH RIVER, THE HS "UZHHOROD"

Data samples (years)	**Probability of exceedance P (1/year, %)**	**Design maxima discharge Q (m^3/s)**	
		Inf	***Sup***
1947-1976	1	1995	2095
	0.5	2545	2730
1947-1984	1	1890	1988
	0.5	2332	2490
1947-1992	1	1788	1908
	0.5	2200	2324
Control sample, 1947-1999	1	1738	1805
	0.5	2113	2222

It should be noted the high values of coefficients of determination (R^2) of the dependencies $d_{m,i} = f(T_{r,m,i})$, $d_{m,j} = f(T_{r,m,j})$, and $d_m = f(Q_m)$ when regression modeling: 0.9996, 0.9999, and 0.9997 in the case of the data sample of 1947-1976 years; 0.9997, 09998, and 0.9998 (1947-1984 years); 0.9997, 09998, and 0.9998 (1947-1992 years); 0.9995, 0.9998, and 0.9983 in the case of the control data sample of 1947-1976 years.

It is worth also noting the goodness of fit of the maxima discharges of 1% probability of exceedance obtained by extrapolation of plotting position probabilities using the proposed method to the Extreme value type I distribution (Gumbell type I, EV1) (See Control data sample of 1947-1999 years). The design discharge of 1% probability of exceedance obtained using the EV1 distribution is 1832 m^3/s. The upper bound estimate (sup) of such a discharge using the proposed method and the Fishburn rule is 1805 m^3/s. The relative prediction error is less than 1.5%. The lower bound estimate (inf) of such a discharge using the proposed method and the Fishburn rule is 1738 m^3/s. The relative prediction error is 5.4%. However, it is worth noting the goodness of fit of the maxima discharges of 0.5% probability of exceedance obtained by extrapolation of plotting position probabilities using the proposed method to the Logarithmic Pearson type III distribution (LP3). The design discharge of 0.5% probability of exceedance obtained using the LP3 distribution is 2130 m^3/s. The upper bound estimate (sup) of such a discharge using the proposed method and the Fishburn rule is 2222 m^3/s. The relative prediction error is approximately 4.2%. The lower bound estimate (inf) of such a discharge using the proposed method and the Fishburn rule is 2113 m^3/s. The relative prediction error is less than 0.8%.

V. Some Discussion Remarks and Conclusions

When we forecast extreme (maxima) hydrological characteristics, can epistemic (non-stochastic) uncertainty be a challenge? Yes, it can. It can be a high challenge. However, the multi-model approach may promote revealing epistemic uncertainty and supporting the choice of a better parametric probability distribution.

There is no proper theoretical or another similar justification for choosing an appropriate probability distribution to forecast maxima discharges of floods using observed data. Plotting position formulas provide a non-parametric means to estimate the observed data probability distribution. Using a plotting position formula, a plot of the estimated values from a theoretical parametric probability distribution can be compared with the observed data. It allows a visual examination of the adequacy of the fit provided by alternative parametric probability distributions.

However, there are more than seventeen different plotting position formulas to fit theoretical parametric probability distributions with the observed data. The issue is the choice of an unbiased empirical formula to plot the observed data. Any plotting position formula can be an option for fitting parametric probability distributions. Based on a multi-model approach, the proposed numerically-analytical method may promote a justification for choosing an appropriate parametric probability distribution to forecast maxima discharges of floods using observed data.

The proposed method applies numerical calculations of empirical probabilities using different plotting position formulas and extrapolation of the divergence between the obtained estimates. The design maxima water discharge estimates forecasted by this method are also noteworthy. In terms of forecast accuracy, these estimates do not differ principally from estimates that can be obtained using well-known, traditional parametric probability distributions.

References

[1] R. Maity, “Statistical Methods in Hydrology and Hydroclimatology,” Springer Transactions in Civil and Environmental Engineering, 2018.

[2] K. Okoli, K. Breinl, M. Mazzoleni, and G. Di Baldassarre, “Design Flood Estimation: Exploring the Potentials and Limitations of Two Alternative Approaches,” Water, vol. 11 (4), 729, 2019, pp. 1–11.

[3] V. Korbutiak, D. Stefanyshyn, O. Lahodniuk, and A. Lahodniuk, “The combined approach to solving issues of the flood hazard assessment using water gauge records and spatial data,” Acta Sci. Pol. Architectura, vol. 19 (1), 2020, pp. 111–118.

[4] J. R. Stedinger, R. M. Vogel, and E. Foufoula-Georgia, “Frequency Analysis of Extreme Events”, In Maidment, D.R., Ed., Handbook of Hydrology, McGraw Hill, New York, Chapter 18 (18.1-18.66), 1993.

[5] Directive 2007/60/EC on the assessment and management of flood risks. Official Journal of the EU, L288/27, 2007.

[6] D. Stefanyshyn, “Testing a numerically-analytical method for prediction design maxima discharges of floods using plotting position formulas: the river Uzh case, the “Uzhhorod” gauging station data,” Environmental safety and natural resources, vol. 46 (2), 2023, pp. 138–162.

[7] H. Apel, A. H. Thieken, B. Merz, and G. Blöschl, “A Probabilistic Modelling System for Assessing Flood Risks,” Natural Hazards, vol. 38, 2006, pp. 79–100.

[8] H. Apel, B. Merz, and A. H. Thieken, “Quantification of uncertainties in flood risk assessments,” Int. Journal of River Basin Management, vol. 6 (2), 2008, pp. 149–162.

[9] D. Stefanyshyn, “Probability assessment of the Kyiv reservoir overflow,” Environmental safety and natural resources, 40 (4), 2021, pp. 73–99.

[10] M. Ren, X. He, G. Kan, et all, “A Comparison of Flood Control Standards for Reservoir Engineering for Different Countries,” Water, vol. 9 (3), 152, 2017, pp. 1–13.

[11] C. Cunnane, “Unbiased plotting positions – A review,” Journal of Hydrology, Vol. 37 (3-4), 1978, pp. 205–222.

[12] L. Makkonen, “Plotting Positions in Extreme Value Analysis,” Journal of Applied Meteorology and Climatology, vol. 45, 2006, pp. 334–340.

[13] H. L. Harter, “Another look at plotting positions,” Communications in Statistics – Theory and Methods, vol. 13 (13), 1984, pp. 1613–1633.

[14] C. Brezinski, and M. Redivo-Zaglia, “Extrapolation and Rational Approximation,” The Works of the Main Contributors, Cham: Springer Nature, 2020.

[15] L. J. Savage, “The foundations of statistics,” New York: Wiley, 1954.

[16] P. C. Fishburn, “Utility Theory for Decision Making,” New York: John Wiley & Sons, Inc., 1970.

[17] P. C. Fishburn, “Non-transitive measurable utility for decision under uncertainty,” Journal of Mathematical Economics, vol. 18 (2), 1989, pp. 187–207.

[18] M. J. Kochenderfer, “Decision-making under uncertainty. Theory and Application,” Massachusetts: The MIT Press, 2015.