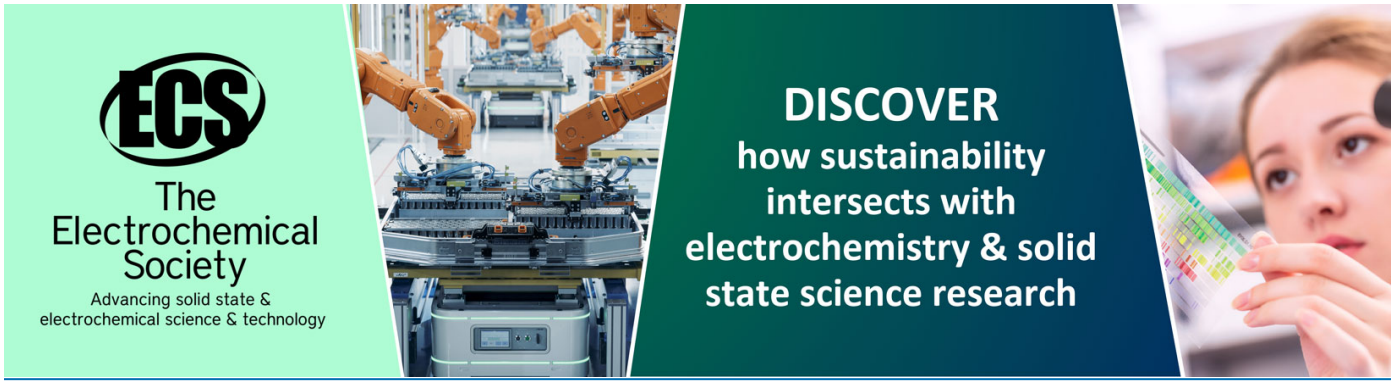
A Comparison of Methods for Calculating Monthly Flows on Small Catchments



A Comparison of Methods for Calculating Monthly Flows on Small Catchments

Milan Cisty[1], Roman Vyleta[1], Veronika Soldanova[1]

[1]Slovak University of Technology in Bratislava, Faculty of Civil Engineering, Radlinskeho 11, Bratislava, Slovakia

milan.cisty@stuba.sk

Abstract. The paper deals with a determination of the monthly time series of flows in catchments where this quantity is not measured. While determining unknown flows, the authors have assumed that the historical climatic data for the given area and flows in nearby and similar catchments are available. So-called "analogy" methods are possible to use for such a defined task. In these methods, unknown flows are determined based on an "analogy" with known data from similar, nearby catchments. The authors have compared hydrological modelling and statistical and machine learning regression methods. They have determined that if at least short-term measurements are available for the river catchment studied, the most suitable method is a regression with LASSO regularisation, as LASSO eliminates the problem of multicollinearity in the input data. The use of the Support Vector Machines and Neural Network with Bayesian regularisation seems to be other suitable methods. The precision of hydrological modelling results is slightly less than the results from regression methods, but the authors have demonstrated that these results are still suitable in the context of water management calculations.

1. Introduction

River flows are essential inputs for various watershed planning tasks and the design of water management constructions. The availability of accurate flows at watersheds provides a basis for various purposes. A continuous time series of river discharges is essential in all stages of water management, including planning, design, and operations but also environmental protection.

This paper deals with the specification of flows in catchments without the direct monitoring of this variable, which is a reality in most of the world's river catchments. For some purposes (e.g., flood protection), daily or hourly time series of flows are required; however, this paper has considered the acquisition of monthly data that are sufficient for different purposes, e.g., the design of irrigation reservoirs or water demands for irrigation.

Regional methods (also known as analogy methods) can be used for this task, as the unknown flows can be computed using an "analogy" with known measured flows and using other data from similar, nearby catchments. In the past two decades, considerable attention in the hydrological literature has focused on this way of modelling flows in water catchments without measurements. Kohnova et al. reported a survey of regional methods used in Slovakia, where the case study presented hereinafter is located, in [1]. An excellent introduction to this topic is available in the work of Hrachowitz et al. [2]. Various types of models have been used for determining unmeasured flows, which fall into four

categories, i.e., physically-based models with distributed parameters (e.g., SWAT, MODFLOW, MIKE SHE), conceptual models with lumped parameters (HBV, TOPMODEL, GR4J, etc.), and various statistical and data-driven models that use regression methods [3] or artificial neural networks [4].

In the present paper, new methods for determining flows on small catchments are evaluated. Four data-mining methods are compared with a hydrological model based on either a daily or monthly time step. The first hydrological model is the conceptual hydrological model, while the second hydrological model utilizes the water balance, which describes the homeostasis between the water input and output from the watershed. The objective is to detect the advantages and disadvantages of these approaches for the given task and develop a methodology for acquiring the monthly flows on streams, where flows are not measured. In section 2., the acquisition and preparation of the data are described. The methods applied in this study are briefly explained in section 3. In section 4., the settings of the experimental computations are described, and the results are evaluated and discussed. Finally, section 5 summarizes the main findings of this study.

2. Case study description

A case study of the Parna stream is reported in this paper. Parna is a small mountain stream in the Small Carpathians in Western Slovakia. Its catchment area is 45.59 km^2. To determine the average daily flow in this stream, known flow data from similar nearby catchments are used (Figure 1), namely, data from water metering stations at Bukova (the Trnavka catchment), Modra (the Vistucky stream catchment), Pila (the Gidra catchment), and Solosnica (the Solosnicky stream catchment).

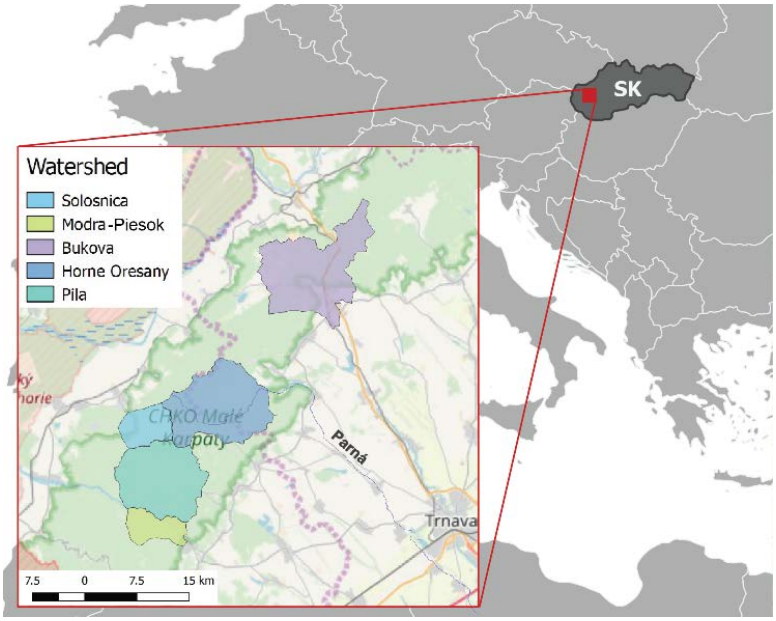


Figure 1. Location of the selected water catchments

Climatic data from the European Climate Assessment & Dataset (ECA&D) were also used in this study. ECA&D contains daily series of observations for 12 elements at 7847 meteorological stations throughout Europe and the Mediterranean, which are provided by national meteorological institutes. The main product of this initiative (E-OBS), which was used in this work, is a daily E-OBS gridded version of the ECA dataset (stored in netCDF) with daily temperature, precipitation and pressure. The E-OBS dataset contains a series of daily observations at meteorological stations throughout Europe and the Mediterranean. The climatic data are provided as a spatial time series for the period 1950–2018; they and have a spatial resolution of 0.25°×0.25°. Data from 1 January 1980 to 31 August 2017 were used. The time series of the daily values of the potential evapotranspiration in the individual water catchments

were calculated using the climatic data. The potential evapotranspiration was calculated using a formula proposed by Oudin in [5].

3. Methods

The limiting factor in the various water management calculations is the fact that the flows in the small streams involved in the study under consideration are, in many cases, not measured. Thus, temporal trends of the flow cannot be evaluated, and both the high and low flows are unknown. Drought or flood protection studies are hard to accomplish without such information. In this paper, regression methods and hydrological modelling are compared for determining such unknown historical time series of flows. For an evaluation of the results, various statistical indicators were used. A brief characterisation of both the regression and hydrological simulation methods used is given below.

An important condition for basic multiple linear regression (MLR) is that the independent variables must not correlate too much; that is, the existence of near-linear relationships between the independent variables (multicollinearity) is inappropriate. Multicollinearity can create inaccurate estimates of the regression coefficients, increase the standard errors of the regression coefficients, decrease the partial t-tests for the regression coefficients, give false or insignificant p-values, and degrade the predictability of the model. However, a quite high level of correlation is likely to occur in the task addressed in this study. The least absolute shrinkage and selection operator (LASSO) was applied in this paper; it redefines the linear regression to prevent the effect of multicollinearity and help ensure a more stable model [6].

In this work, two machine-learning algorithms were applied, namely, Support Vector Machines and the Bayesian Regularized Feed-Forward Neural Network. They were used for the supervised learning problem in this study, where we used the training data (with multiple features) to predict a target variable.

A Support Vector Machine (SVM) [7] is an effective, supervised machine learning method with the possibility of also using it for regression tasks. It is specific by using the kernel trick, i.e., nonlinear mapping is used to transform the original training data of a nonlinear problem into a higher dimension. SVM learns a nonlinear function indirectly, i.e., it learns a linear function in the space induced by the kernel, which matches a nonlinear function in the original space, where it is mapped backwards. The next, most important, concept in SVM methodology is that it ignores small errors. In SVM, the threshold for this simplification is set by defining a loss function that ignores errors that are situated within the distance ε (a tuned parameter). This type of function is called an epsilon-insensitive loss function. As a consequence, a good generalization of SVM is gained, because not all the data vectors are used while building the model, but only the so-called support vectors (i.e., the noise is removed.

Artificial Neural Networks (ANN) has been described in numerous previous works, so only a limited explanation follows. Briefly summarized, a neural network consists of input, hidden, and output layers, all containing some amount of neurons. The number of neurons in the input layer and the output layers corresponds to the number of input and output variables of the model; the number of neurons in the hidden layer is the tuned parameter. In ANN training overfitting can occur; this means that it works well on training data, but can have an unsatisfactory performance on previously unseen testing data. Regularization (shrinkage) in ANN permits improvement of this situation. Bayesian regularization is an effective regularization technique that is similar to that used in LASSO regression. The Bayesian Regularized Feed-Forward Neural Networks (BRNN) model was included in the computational methods used because overtraining was observed while modelling with more complex neural networks.

Two hydrological models were also used, one for the daily time step and one for the monthly time step. The lumped conceptual rainfall-runoff hydrological model (TUW model) [8] runs on a daily time

step and consists of snow, soil moisture, and flow routing routines. The snow routine simulates snow accumulation and melting using a degree-day concept. The soil moisture routine simulates runoff generation and changes in the soil moisture state of a catchment. The upper and lower soil reservoirs represent quick and slow runoff routing. A genetic algorithm was used to calibrate the 15 parameters of this conceptual model.

Using the second model, the authors of this paper tested whether it is more appropriate to calculate monthly flow rates directly and not their daily values initially, which are subsequently converted to monthly values as in the previous case. A refinement of the WatBal model [9, 10] was used for estimating the mean monthly discharges with the scheme shown in Figure 2.

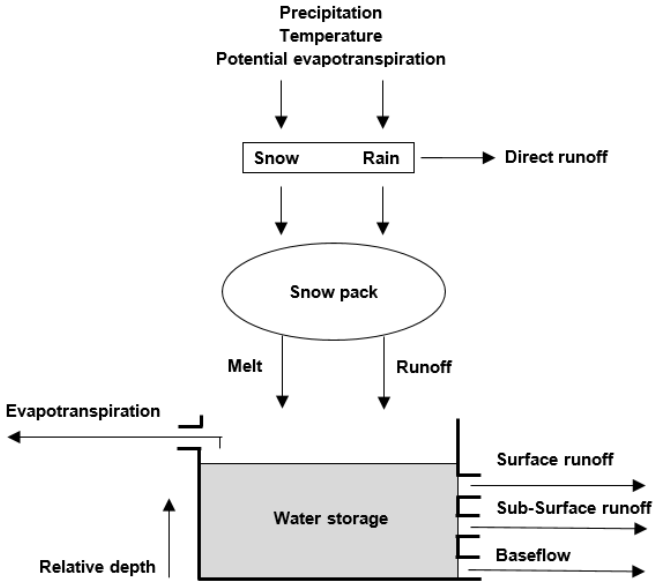


Figure 2. Structure of the conceptual HyBal hydrological balance model

The inputs to the Hydrological Balance (HyBal) model include the mean monthly precipitation, the mean monthly air temperature, the mean monthly discharges at the catchment outlet, and the mean monthly potential evapotranspiration. The model schematizes the river basin by dividing it into two nonlinear reservoirs: in the first nonlinear tank, the process of the accumulation and melting of snow takes place; in the second nonlinear tank, the simulation of the water balance of the catchment's hydrological processes takes place. The underlying assumption of the model is that the individual components of the runoff from the basin depend on the actual volume of water in the watershed.
The mass balance equation in the model is written as:

$$S_{(i)} - S_{(i-1)} = \left[\left(R_{eff\,(i)} \left(1-\beta\right) \right) - R_{s(i)} - R_{ss(i)} - E_{a(i)} - R_b \right] \Delta t \qquad (1)$$

where: $S_{(i)}$, $S_{(i-1)}$ is the water currently stored in the basin in months i and i *-1* [mm]; i is the time step [month]; $R_{eff\,(i)}$ is the effective precipitation in the month i [mm]; β is the direct runoff coefficient [–]; $R_{s(i)}$ is the surface runoff in the month i [mm]; $R_{ss(i)}$ is the subsurface runoff in the month i [mm]; $E_{a(i)}$ is the basin's average actual evapotranspiration in the month i [mm]; and R_b is the base flow [mm]

4. Results and discussions

4.1. Selection of suitable river catchments

The river basins for the analogous calculations should be similar to the river catchment where the flow is intended to be determined. However, the flow characteristics of the streams cannot be compared

directly, as they are unknown for the target stream. Only the various physical characteristics of the watershed can be compared, as they influence the outflow.

The outflow regime of a river catchment depends on its climate conditions and topographic features such as its altitude, slopes, exposition, the density of the drainage network, etc. Also, the geology types of prevailing soil in the watershed, and land use of the area, e.g., whether forests, meadows, arable soil or urbanized areas occur in the catchment, are important. Some climate, topological, geological, and other properties change relatively smoothly, so it is helpful to choose nearby watersheds if they have been measured (Figure 1). Several analyses have been developed for this purpose; the selected results are briefly presented in Figure 3 and Tables 1 and 2.

Figure 3 presents a comparison of the soil types and land use in the individual river catchments. The soil types differ according to the content of sand, silt, and clay particles in the soil. Different soil types vary in terms of the ratio of infiltration and the outflow of water during periods of rain, in the ability to retain water in the soil, and in other properties that influence the outflow of water from the river catchment. Similarly, land use (for example, a forest versus arable land) also has a significant impact on river basin drainage properties.

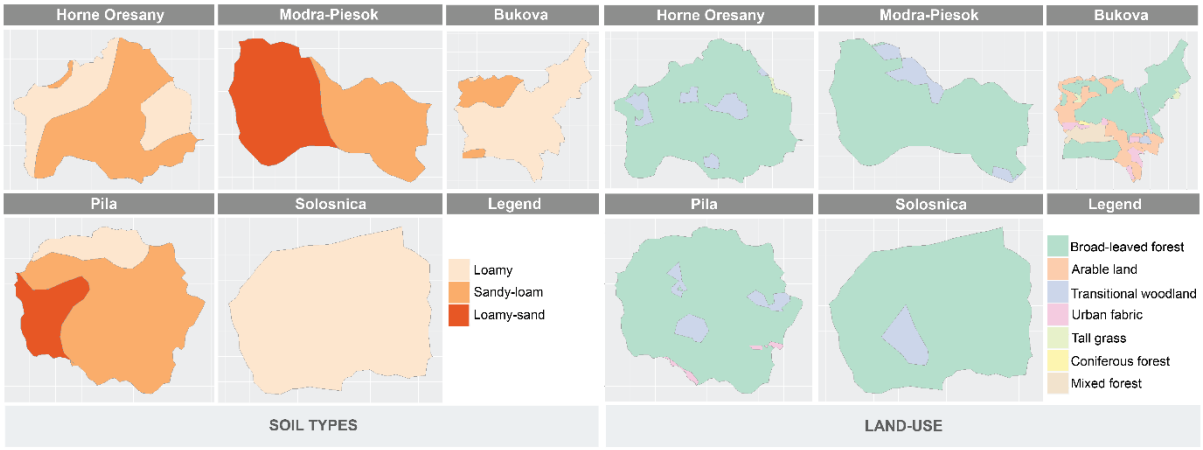


Figure 3. Soil types and land use in the river catchments addressed

The representative part of the analyses is evaluated in tables 1 and 2. These tables and the GIS analyses show that for the river catchments assessed and from the point of view of the catchment features influencing the outflow regime, the Parna river catchment is most similar to the Gidra catchment (Pila gauging station) and the Vistucky stream catchment (Modra-Piesok station). Therefore, these two catchments will be preferred in the following calculations.

Table 1. Geographic characteristics of the watersheds

Watershed	Area (km^2)	Surface Characteristics (median)			Soil Types (%)			Land Use (%)				
		Elevation (m a.s.l.)	Slope (°)	Aspect (°)	Loamy	Loamy-sand	Sandy-loam	Arable land	Broad-leaved forest	Mixed forest	Transitional woodland/shrub	Urban fabric
Horne Oresany	37.3	403.1	11.1	151	38	0	62	0	90.9	0.8	8.2	0
Bukova	43.0	332.4	9.0	170	82.5	0	17.5	23.9	60.7	6.9	2.1	4.9
Modra-Piesok	9.4	495.1	8.0	92.9	0	54.3	45.7	0	91.5	0	8.5	0
Pila	32.9	426.7	9.8	161	12.8	16.3	70.9	0	92.4	0	7.1	0.5
Solosnica	10.5	420.8	16.3	195	100	0	0	0	94.4	0	5.6	0

Table 2. Hydrological characteristics of watersheds

Watershed	Calibration ability (NSE)	Hydrological network characteristics		Morphometric characteristics			Topographic wetness index		
		boundary segmentation	Form factor	First-order stream length (km)	Total channel length (km)	Drainage density (km/km^2)	Min	Max	Mean
Horne Oresany		1.3	0.15	17.1	56.4	1.51	-12.0	11.2	4.3
Bukova		1.7	0.18	25.7	48.4	1.13	-10.0	12.3	4.7
Modra-Piesok		1.4	0.11	6.0	14.2	1.51	-9.0	11.2	4.4
Pila		1.2	0.27	16.5	54.7	1.66	-8.4	11.8	4.5
Solosnica		1.2	0.23	4.0	15.8	1.51	-9.7	7.4	3.8

4.2. Hydrological modelling

The calculations with the TUW model were performed in daily steps, and the daily flows were subsequently converted into monthly flows. The calibration was implemented based on the flow and climate data from the Pila catchment, which was assessed as being the most similar to the river catchment in which the unknown flows were to be calculated. The optimal parameter values of the TUW model were acquired by a genetic algorithm using the flow and climate data from the Pila catchment. These parameters were subsequently applied in the modelling of the river catchment with the unknown flows (Parna stream) using climate data from the Parna watershed. The genetic algorithm population was set at 500, the number of parameters to be determined was 15, and the maximum number of generations was 20. The objective function minimises the Nash–Sutcliffe efficiency, which is often used in hydrological modelling [11]. The calculated daily and monthly flows on the Parna stream are compared with the measured values in Table 3.

In the calibration procedure of the HyBal hydrological balance model, 11 model parameters are optimized. As the investigated catchment is not gauged, the traditional approach when the water balance model is calibrated against the streamflow data at the catchment outlet was not possible. Within this study, the following approach was applied. In the first step, a gauged catchment of the River Gidra at Pila was selected as a catchment with similar physiographic characteristics as the ungauged catchment.

In the next step, the model was calibrated, and its parameters were transferred to the ungauged catchment and used to simulate a time series of mean monthly discharges of the same length using the climatic inputs related to the ungauged catchment.

4.3. Statistical and machine learning modelling

In regression calculations, climate and flow data from surrounding river catchments were used as independent variables. For determining the unknown historical flows, the assumption was made that the measurement of the flow also recently started in the Parna river catchment. Only the remaining flows at Parna from the period 1980 - 2017 should therefore be computed, and that short period of measurements will be used when creating a regression model. It is assumed that the measurements started at the beginning of 2016. Based on this period, the regression models were derived and consequently applied to the whole historical period of interest in which the flows were to be calculated.

The statistical and machine learning methods described in the previous subsections were used for the regression calculations. The calculations were performed in R [12] using the same data as was used in the hydrological modelling, but some feature engineering was made with them.

As the flow from a catchment is influenced not only by the current values of the climate variables but also by their values from previous days, we also included climate data from seven days before the date of the prediction. The previous history of the hydro-climatic conditions is described by three variables summarizing the past precipitation (cumRAIN7, cumRAIN14, cumRAIN21) and evapotranspiration (cumPET7, cumPET14, cumPET21). The numbers in these variable names denote how many days were summarised backwards. In this way, a training set with 35 explanatory variables was created. Since this set has quite a lot of explanatory variables, the most suitable ones among them were chosen to avoid overtraining of the models. Based on the training file, its 1000 variations were created by bootstrapping, and a linear model with four explanatory variables was found for each variation. The most frequent variables included in these models are shown in Figure 4a. Figure 4b also shows the variable importance in computations but takes into account the interaction of the variables. Based on this analysis, ten variables were selected for the training set, which was used in all the regression calculations, namely, the first nine variables of Figure 4a and the fourth to sixth variables from Figure 4b. This set covers the period of anticipated short-term measurements on the Parna stream for 608 days (only data up to August were available for 2017). The test file includes the same variables, but the data relates to the whole period of 1980–2017, i.e., it contains 13,738 lines (one per day).

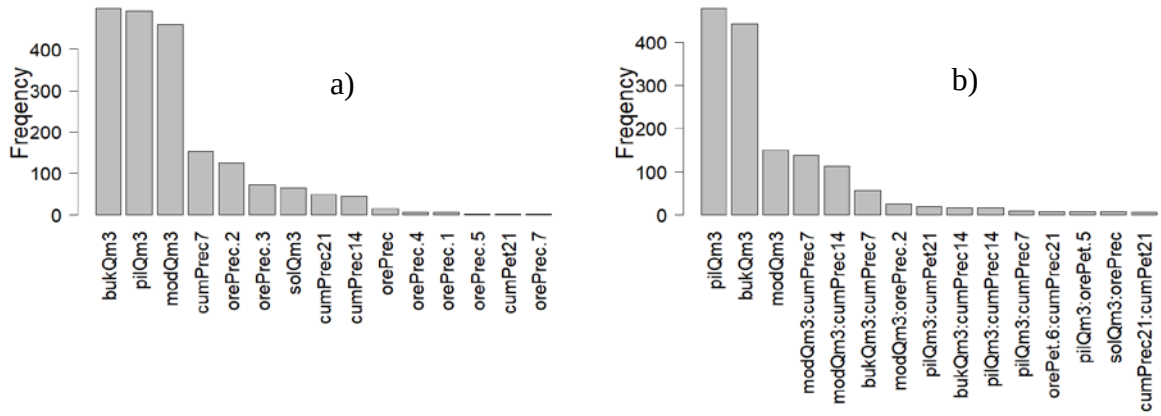


Figure 4. Important variables – a) individual variables, b) possibility of variable interactions included (the variables are separated by a colon). The abbreviations in variable names mean: Qm^3 – flow, Prec – precipitation, Pet – evapotranspiration, cum – cumulative variable, 2 – two days before the prediction of the flow, etc.

The results of the regression and hydrological calculations are presented in table 3.

Table 3. Final evaluation of the models

model	ME	RMSE	PBIAS %	NSE	R2
TUW model	-0.021	0.175	-6.5	0.651	0.659
HyBal	-0.030	0.191	-9.0	0.586	0.603
MLR	0.013	0.146	4.0	0.758	0.768
LASSO	0.017	0.140	5.1	0.776	0.781
BRNN	0.004	0.144	1.1	0.764	0.764
SVM	-0.012	0.143	-3.8	0.766	0.775

ME – Mean Error, RMSE – Root Mean Square Error, PBIAS% - Percentual Bias, NSE - Nash-Sutcliffe efficiency, R2 – coefficient of determination

Table 3 shows that if at least short-term measurements are available for the river catchment studied, the most suitable method for determining the unmeasured flows is a regression with a LASSO regularisation. The Support Vector Machines and Neural network with Bayesian regularization also seem to be other appropriate methods. However, table 3 indicates that similar results can also be obtained using other statistical and machine learning methods. However, there is a question as to how the hydro-climatic conditions during the calibration period will affect the possibility of using the measurements accomplished in this period for the calibration. Figure 5 shows a histogram of the NSE values, i.e., the precision of computing flows when all the possible one and two-year periods from the complete investigated period of 1980 - 2017 were considered as the calibration period. The chart shows that fewer than 10% of these periods lead to unsatisfactory results (the NSE is less than 0.5). Based on further analyses of these results, it can be seen that in the vast majority of cases, these are the same periods based on which it is also not possible to calibrate the Pila watershed. The possibility of calibrating the Pila watershed can be verified because the flows in the Pila watershed were measured during the entire historical period investigated (the flows in the Pila are the main explanatory variables - Figure 4). In this reciprocal manner, an unsuitable period can be identified, i.e., a period for which the measurements are not proper for calibrating the model, so another method for determining the flows should be searched for (i.e., not a regression).

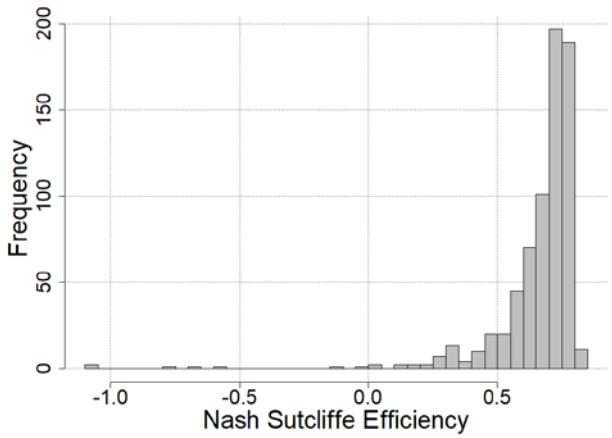


Figure 5. Histogram of the NSE values obtained when all the possible one and two-year periods from the complete investigated period of 1980 - 2017 were considered as the calibration period

When the measurements are not suitable for calibrating a regression model or are not performed at all, a hydrological model can be used for the given task. Modelling with a daily time step (the TUW model) gives better results than the model with a monthly time step (HyBal), because due to a shorter time step, it has the opportunity to the better capture hydrological processes in the watershed. The precision of the hydrological modelling results is slightly less accurate than the results from the regression methods. However, in Figure 6, which shows the observed and computed flows, we can see that the inaccuracies that are more significant mostly occur in March. If the accuracy of the flows in March is not particularly important (for example, when designing irrigation), the results of the hydrological modelling are satisfactory. Otherwise, for this month, an exclusive model may be calibrated.

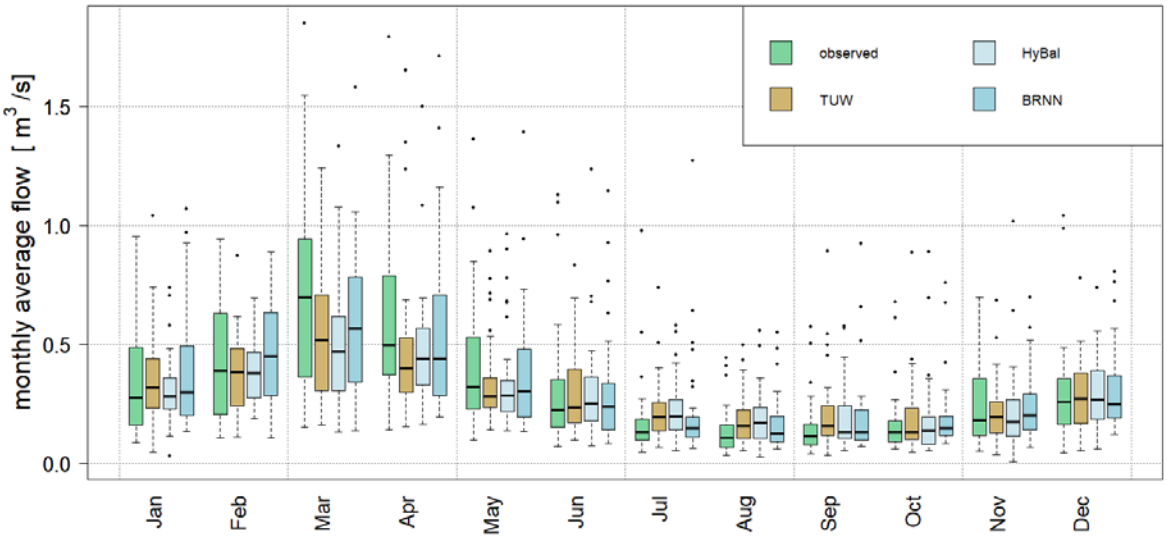


Figure 6. Summary of the selected flow modelling results by months

5. Conclusions

The objective of this study was to compare various methods of calculating stream flows on ungauged catchments. The methods examined include regression and hydrological modelling methods.

If at least short-term measurements are available for the relevant river catchment, the most suitable way evaluated in this study is the flow calculation method using regression with LASSO regularisation. Support Vector Machines and a neural network with Bayesian regularization seem to be the next suitable methods.

If no flow data are available, a hydrological model can be used. The accuracy of modelling smaller streams by this type of model is limited, as various irregular phenomena (water spurts, water leaking to the underground, etc.) affect catchment hydrology relatively more than in the case of larger river basins. For this reason, we even considered the results of hydrological models, although they were evaluated in this study as less accurate in comparison with regression, as satisfactory (Figure 6). In this paper, two conceptual hydrological models were applied. A model with a daily time step (the TUW model) gives better results than a model with a monthly time step (HyBal) because due to the shorter time step, the TUW model can better capture hydrological processes in the watershed. The results of the hydrological modelling were less precise than those from the regression methods, but they can still be applicable within the context of the engineering calculations.

Acknowledgements
This work was supported by the Slovak Research and Development Agency under Contract No. APVV-15-0489 and by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences, Grant No. 1/0662/19.

References

[1] S. Kohnova, J. Szolgay, L. Solin, and K. Hlavcova, "Regional methods for prediction in ungauged basins," *Ostrava: Key Publishing*, p. 113, 2006, ISBN: 80-87071-02-6.
[2] M. Hrachowitz, H. H. G. Savenije, G. Blöschl, J. J. McDonnell, M. Sivapalan, J. W. Pomeroy, and F. Fenicia, "A decade of Predictions in Ungauged Basins (PUB)—a review," *Hydrological sciences journal*, vol. 58(6), pp. 1198-1255, 2013.
[3] M. S. Gibbs, H. R. Maier, and G. C. Dandy, "A generic framework for regression regionalization in ungauged catchments," *Environ. Model. Softw.*, vol. 27, pp. 1-14, 2012.
[4] V. Kuzmin, I. Pivovarova, K. Shemanaev, D. Sokolova, A. Batyrov, A. N. Tran, and D. Dang, "Method of Prediction of the Stream Flows in Poorly Gauged and Ungauged Basins," *Journal of Ecological Engineering*, vol. 20(1), pp. 180-187, 2019.
[5] L. Oudin, C. Michel, and F. Anctil, "Which potential evapotranspiration input for a lumped rainfall-runoff model?: Part 1 - Can rainfall-runoff models effectively handle detailed potential evapotranspiration inputs?," *Journal of Hydrology*, vol. 303/ Nos. 1-4, pp. 275-289, 2005.
[6] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning: data mining, inference, and prediction," *Springer Series in Statistics*, 2009.
[7] V. Vapnik, "The nature of statistical learning theory," *Springer-Verlag: New York*, 1995.
[8] A. Viglione, and J. Parajka, "TUWmodel: Lumped Hydrological Model for Education Purposes," R package version 1.0-1., 2018 [Online] Available at: https://CRAN.R-project.org/package=TUWmodel.
[9] D. Yates, "WatBal: an integrated water balance model for climate impact assessment of river basin runoff," *International Journal of Water Resources Development*, vol 12(2), pp. 121-140, 1996.
[10] R. Výleta, J. Szolgay, and K. Hlavčová, "Calibration of rainfall-runoff balance model with monthly time step for the upper Hron river," *Acta Hydrologíca Slovaca*. vol. 10, no. 2, 213-220, pp. 183-190, 2009, ISSN 1335-6291 (in Slovak).
[11] D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith, "Model evaluation guidelines for systematic quantification of accuracy in watershed simulations," *Transactions of the ASABE*, vol. 50(3), pp. 885-900, 2007.
[12] R Core Team, "R: A language and environment for statistical computing," *R Foundation for Statistical Computing*, 2018 [Online] Available at: https://www.R-project.org.