

# Base Flow Index Estimation on Gauged and Ungauged Catchments in Hungary Using Digital Filter, Multiple Linear Regression and Artificial Neural Networks

Zsolt Jolánkai<sup>1\*</sup>, László Koncsos<sup>1</sup>

RESEARCH ARTICLE

Received 14 February 2017; Revised 02 October 2017; Accepted 14 November 2017

## Abstract

*A country scale analysis of diffuse source nutrient emissions have been undertaken previously on small catchments level using the MONERIS model, which needed a proper estimation of surface and subsurface runoff differentiation to support or contradict its own water budget based method. As reliable, country scale base flow estimation has not been available for the country at the time of the study, this knowledge gap has been tried to be filled by the current work. This has been done using multiple methods. Digital filter have been applied on continuous river discharge data on gauged catchments in order to determine base flow indexes (BFI). Using multiple linear regression (MLR) and artificial neural networks (ANN), climatic, soil and land use properties of the catchments have been used to extend base flow indexes to ungauged catchments. MLR brought acceptable results (adjusted  $r^2$  values around 0.7), however it proved to be sensitive of the selection of catchments used for validation, and therefore a mean of prediction by thirty different regression equation was used for the estimation. ANN was less sensitive for the change of the variables included and the number of nodes used for the learning. The results are comparable with the MLR method and show good agreement in most of the areas, however in some part of the country the two approach show significant differences in the predicted BFI values.*

## Keywords

*moneris, base flow index, multiple linear regression, artificial neural network, base flow separation, digital filter*

## 1 Introduction

The current study is undertaken to support the work of country scale estimation of diffuse source nutrient emissions to the surface water bodies of Hungary, using the Moneris model [1]. To access the nutrient emissions properly, the ratio between surface runoff and subsurface runoff is essential, as these pathways deliver nutrients in significantly different concentrations. Catchment scale and country scale estimation of the base flow indexes (BFI, ratio of the subsurface flow compared to the total runoff in a water body) has been done in Hungary by different experts, but the methodology and the results have not been published, therefore the comparison with the current study cannot be delivered. The current study aims to regionalise the base flow indexes to ungauged areas to provide a control of existing water balance estimations or a base for a new water balance estimation.

Hungary is situated in the middle of the Danube basin, and in the middle of the Carpathian basin. The large mountains surrounding this basin lie outside of the borders, only low altitude mountains (mostly below 1000 m above Baltic Sea Datum) can be found in the north-eastern and mid-western part of the country.

West from the Danube, the so called Trans-Danubian part is primarily hilly, with karstic limestone features in the Trans-Danubian Mountains. The hilly part have cambisols in most parts over loess-like deposits [2]. The flat areas east of the Lake – Balaton and in the very north-western part are very fertile regions with loamy chernozem soils over fluvial deposits [2]. The area between the Danube and the Tisza River is called the Danube-Tisza Interfluve that drains to the Danube on one side and to the Tisza in the other side. It is characterised by small altitude differences, semi-arid climate and primarily sandy soils [2]. The eastern part of the country is part of the Tisza basin, and is predominantly flat (except from the Northern Mountains, and the slightly hillier sandy Nyírség) and the soils and deeper geological layers are characterised by fine texture fluvial sediments. The Northern Mountains have volcanic formations and karstic mountains as well. The flat area is called the Great-Plain (Alföld), which belong to a 100

<sup>1</sup> Department of Sanitary and Environmental Engineering,  
Faculty of Civil Engineering,  
Budapest University of Technology and Economics  
H-1111 Budapest, Muegyetem rkp 1-3, Hungary

\* Corresponding author, email: [jszolt@vkkt.bme.hu](mailto:jszolt@vkkt.bme.hu)

000 km<sup>2</sup> large geographic unit, of which 52 000 km<sup>2</sup> lie within the borders of Hungary [3] and the total altitude difference is approx. 100 meters, but the average relief is only a few meters within the most part of the area. A large part of the Alföld is an artificially drained area, with a total length of approximately 40 000 km of open drainage channels [4].

The largest river is the Danube, and the largest river basin belong to the Tisza River, which is the largest tributary river of the Danube. The river network of the country contains ca. 9800 river sections, but over 90% of the total discharge is carried by the rivers coming from outside the borders of the country.

Due to its geographic location the country is characterised by shallow groundwater (2–5 m on average [5]).

The annual precipitation in Hungary spread between 500 (Alföld region) to 900 mm (most western parts), while the actual evapotranspiration spreads in the same range [5]. At the Alföld area, the evapotranspiration exceeds the precipitation in many years.

While in the mountainous and hilly area of the country, the river stage/discharge is measured on daily bases, the small channels on the flat areas of the country are ungauged. The flows are difficult to be measured, because the artificial channels are often unable to drain gravitationally and are pumped when the receiving water bodies have high water stages.

## 2 Methodology

For the current analysis four distinctive steps have been taken: river discharge data have been analysed by digital filter based hydrograph separation technique on selected gauged catchments and base flow indices have been calculated for these catchments; catchment properties have been determined by GIS based calculation; BFI values have been correlated to catchment properties by simple and multiple linear regression analysis and BFI values have been predicted on ungauged catchments; artificial neural networks have been built on the training dataset of catchment properties and BFI predictions have been made on ungauged catchments.

### 2.1 Data availability

The river discharge time series has been provided by the National Water Directorate. The Discharge monitoring network consists of 234 stations. The longest time series in the database are dating back to 1950, but there are new stations with only a few years of data. Data with regard the base flow indexes were not available at the time of the analysis.

The catchment property database is based on the database built for the MONERIS nutrient emission estimation model during the RBMP revision in 2015 [6]. The data used for the current analysis can be found in Table 1. Note that most calculations in the table were based on data received from the National Water Directorate (OVF) and the data is unpublished.

### 2.2 Selection of sites for base flow index estimation

The selection criteria for the monitoring stations included in base flow separation have been multiple: the corresponding catchment size was the first filter, set between 0 and 1000 km<sup>2</sup> (earlier the lower limit was 100 km<sup>2</sup> as found in earlier research [9] but as there are many smaller catchments where prediction was to be made, it was dismissed), the other criteria was the length of continuous time series. A minimum length of 365 day of continuous discharge measurement have been selected as a threshold as yearly average BFI was aimed to be determined, which needed the inclusion of seasonal differences in the flow regimes. Another criteria for the selection was that stations where any part of the catchment belong to an upstream neighbouring country has to be closed out from the selection, as catchment properties were unknown for such catchments. Stations, where flow is significantly influenced by the artificial water intake from other water bodies also had to be dismissed.

The time frame for the time series selection has been set to 2001 to 2010, a relatively short period. The reason behind this is that the estimation of the diffuse source pollution is 2009 to 2012, therefore older data would not have been reasonable to be included and most of the time series are only up to 2010, the rest have not been registered yet in the national database. The other reason is that many station with valuable data have been commissioned after 2000.

Altogether 87 stations have been selected for base flow separation (Fig. 1), and further 23 for under 100 km<sup>2</sup> catchment size. Lowland areas of the country misses flow data, therefore the regionalisation of the baseflow index values are expected to be most uncertain on this area.

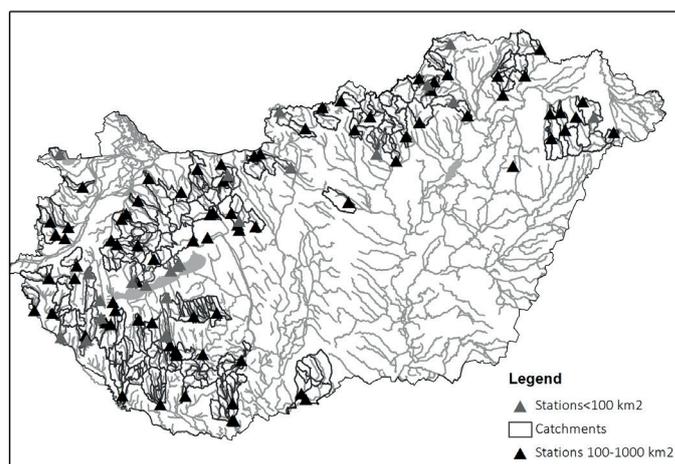


Fig. 1 Discharge monitoring station below 100 and between 100 and 1000 km<sup>2</sup>

### 2.3 Base flow separation

The base flow was separated from the total measured channel flow using a digital filter method suggested by Arnold [10], (also referred to as BFLOW). The reason for selecting this method, is because large number of digitally available time series had to be processed, and the technique referred to is automated, therefore

is easily processable by computational programmes, which have been used for the current study as well. The automated technique used in this research is also widely used in base flow separation studies [11] and has been assessed in some review studies, supporting its validity [12, 13], the latter even suggesting that from some aspects, the BFLOW technique simulates base flow behaviour the best. All review studies warn however, that base flow behaviour is complex and therefore simple filter methods are not too accurate, however no better method is currently available to separate base-flow from surface runoff.

The original separation algorithm used in the above study was developed by Nathan and McMahon [14], using the following equation to estimate the quick response runoff (1):

$$q_t = \beta \cdot q_{t-1} + \frac{(1+\beta)}{2} \cdot (Q_t - Q_{t-1}) \quad (1)$$

Where  $q_t$  is the filtered runoff at the time step  $t$  (one day),  $Q$  is the original streamflow, and  $\beta$  is a filter parameter, which is set to 0.925 according to Arnold et al. [15]. From  $q_t$ , baseflow can be calculated according to (2):

$$b_t = Q_t - q_t \quad (2)$$

According to above cited study, the filter can be used many times (forward, backward and forward) on the flow time series to achieve a smoother base flow curve, however the most realistic base flow index is given in the region somewhere between the first and second pass, when compared with other methods [10]. For the current study this practise has been followed and the final base flow index value was calculated as the average value between the first and the second pass of the filter. In lack of the previous application of other separation methods on the flow time series on Hungarian flow data, the results could not be compared, only visual validation (assessment of the flow hydrographs according to the graphical base flow separation techniques [16]) has been carried out on randomly selected flow data. Further studies on base flow estimation are required (tracer tests, multi-stage measurements, visual separation etc.) to provide comparison of the current results.

## 2.4 Selection of catchment properties for BFI estimation using MLR and ANN

A large number of climatic, physiographic and land use controls have been calculated for each catchments belong to the selected flow monitoring stations (Table 1). Underlying geological formations have been the most untrusted data, as national permeability map has not been available. Soil sand content and clay content have been calculated from national soil database using the methodology described in the annex 3.1 of the national river basin management plan (RBMP) 2015 [6]. The data used for this work is entirely based on the database collected for the diffuse source pollution load estimation done for the RBMP revision in 2015 [16].

The selection of the most appropriate variables for the description of the base flow index was based on multiple regression analysis best model method as described below. After the first trials further filters have been used for catchments selection for BFI calculations. Catchments, where the proportion of the areas with good porosity carbonate rock is more than 10 % of the total catchment size (expert estimate), have been filtered out, which caused a significant increase in the estimation accuracy. Furthermore, catchments, where the proportion of the point source discharge is more than 10 % of the total discharge (expert estimate) have been discarded from the further analysis. After this procedure, 64 catchments remained for BFI analysis.

## 2.5 Multiple linear regression analysis

A multiple linear regression algorithm has been used to find best fitting model for BFI estimation based on catchment properties as earlier have been done by several authors [9, 17, 18, 11, 19]. The optimisation procedure was using least square method as objective function, to maximise correlation, with a hypothesis that the errors have standard normal distribution and are independent. The overall fitness of the resulting equation is assessed by adjusted  $R$  squares (3).

$$R_{adj}^2 = 1 - \left(1 - R^2\right) \cdot \frac{n-1}{n-p} \quad (3)$$

Where  $R^2$  is the square of Pearson correlation,  $n$  is the number of the samples and  $p$  is the number of explanatory variables used for the regression. The number of variables is also a variable during the optimisation process, which results in a best fitting model, with the highest adjusted  $R^2$  value. Out of the total 64 catchments, 54 was used for regression, and 10 was used for validation. The latter have been selected randomly. As suggested by [20] the linear regressions used for BFI regionalisation are not very robust, as found in many studies, hence the best fitting models have been searched thirty times in the current study with different selection of the validation catchments in order to get information on the uncertainty related to the estimation and instead of using the equation with the highest  $R^2$  value, an average of the model runs have been used for prediction. Apart from best model approach, forward and backward stepwise multiple linear regression analysis [21] have been used for comparison.

Calculation of uncertainty of the mean from the multiple BFI value estimates for ungauged catchments were based on the following equation (4):

$$\Delta BFI_{avg} = \frac{\sigma}{\sqrt{N}} \quad (4)$$

Where  $\Delta BFI_{avg}$  is the uncertainty related to the average of the  $BFI$  estimation values,  $\sigma$  is the standard deviation, and  $N$  is the number of trials (samples).

**Table 1** Catchment properties used for the regionalisation of the BFI values

Catchment property	Description	Reference
Catchment area	Total size (km <sup>2</sup> )	GIS data for RBMP 2015 (Source: National Water Directorate, unpublished)
Grassland	In % of the catchment size	Own calculations based on Corine Land Cover [7]
Urban areas	In % of the catchment size	Own calculations based on Corine Land Cover [7]
Woodland, schrubland	In % of the catchment size	Own calculations based on Corine Land Cover [7]
Total agricultural land	In % of the catchment size	Own calculations based on Corine Land Cover [7]
Unconsolidated rock areas near groundwater	In % of the catchment size	Own calculations based on Agrotopo database (RIISAC 1991) and groundwater table depth map (RBMP1, 2009, unpublished)
Unconsolidated rock areas far groundwater	In % of the catchment size	Own calculations based on Agrotopo database (RIISAC 1991) and groundwater table depth map (RBMP1, 2009, unpublished)
Solid rock areas with good porosity	In % of the catchment size	Own calculations based on Agrotopo database [2]
Solid rock areas with poor porosity	In % of the catchment size	Own calculations based on Agrotopo database [2]
Clay content	Average clay content in the catchment (%)	[6]
Sand content	Average sand content in the catchment (%)	[6]
Average elevation	Average elevation of the catchment (mAOD)	Own calculation based on 25 × 25m DEM raster (Source: National Water Directorate, unpublished)
Average slope	Average slope of the catchment (%)	Own calculation based on 25 × 25m DEM raster (Source: National Water Directorate, unpublished)
Flat areas	Percent of the total catchment size, which has a slope less than 1 %	Own calculation based on 25 × 25m DEM raster (Source: National Water Directorate, Unpublished)
Average evapotranspiration	Average evapotranspiration in the catchment between 2000–2009 (mm)	Calculations based on country scale evapotranspiration map [8]
Average temperature	Average temperature in the catchment between 2000–2009 (°C)	Calculations based on pointwise meteorological data (Source: National Water Directorate, unpublished) More at [6]
Average precipitation	Average yearly precipitation in the catchment between 2000–2009 (mm)	Calculations based on pointwise meteorological data (Source: National Water Directorate, unpublished), more at [6]
Length of river reaches	Total length of channel reaches in the catchment as per data base (km)	Calculations based on national river dataset (Source: National Water Directorate, unpublished)
Sandy agricultural soils	Average % of total agric. area	Derived from Agrotopo [2] database
Loamy agricultural soils	Average % of total agric. area	Derived from Agrotopo [2] database
Silty agricultural soils	Average % of total agric. area	Derived from Agrotopo [2] database
Clayey agricultural soils	Average % of total agric. area	Derived from Agrotopo [2] database
Groundwater recharge area	Average % of total area	Unpublished data from RBMP 1 (2009)
Flow from point source emissions	m <sup>3</sup> /sec	Calculations based on TESZIR database (Source: NWD [6])
Flat lowland areas	% of total area	Own calculations based on DEM
Flat highland areas	% of total area	Own calculations based on DEM
Precip-ET	Difference between estimated average yearly precipitation and evapotranspiration	Calculation based on Prec and ET maps (see above)
Limestone fraction	% of limestone of total area	Derived from Agrotopo [2] database
Andesit, basalt fraction	% of andesit+basalt of total area	Derived from Agrotopo [2] database
Granite fraction	% of granite of total area	Derived from Agrotopo [2] database
Tertiary deposits fraction	% of tertiary dep. of total area	Derived from Agrotopo [2] database
Clay-slate fraction	% of clay-slate of total area	Derived from Agrotopo [2] database
Wetland fraction	% of wetlands of total area	Derived from Corine Land Use map [7]
Lake surface areas on tributaries	Total lake surface area (km <sup>2</sup> )	Derived from GIS data for RBMP 2015 (Source: NWD)
Lake surface areas on main channel	Total lake surface area (km <sup>2</sup> )	Derived from GIS data for RBMP 2015 (Source: NWD)
Lake surface area on catchment outlet	Total lake surface area (km <sup>2</sup> )	Derived from GIS data for RBMP 2015 (Source: NWD)
Total channel length	Length of channels (natural and man made) within the catchment (km)	Derived from GIS data for RBMP 2015 (Source: NWD)

## 2.6 Artificial Neural Networks (ANN)

The use of ANN-s in hydrology has quite a long record, and it has been also used by some authors in BFI regionalisation [20, 17, 22]. It has been found in the cited studies that ANN-s perform well in describing the underlying processes in a black-box manner. The disadvantage of this approach is that the processes cannot be understood as clearly as from linear equations, but as natural processes in catchment hydrology are mostly non-linear [23] the linear equations might not work at all.

For the current work a feed forward, backpropagation [24], multi-layer neural network system (MATLAB) have been used. The model have been used with one hidden layer and the number of nodes have been varied throughout the analysis to see the sensitivity of the results on the number of nodes. Levenberg-Marquardt optimisation (LMA) algorithm [25] and Bayesian regularisation algorithm (BRANN) [26] have been used. The latter have been used as it was designed to give more robust results in small size, noisy samples and it is difficult to overtrain and overfit [26]. The same catchments have been used for the ANN analysis that for the MLR analysis. Out of the 64 catchments, 57 have been used for training the system and randomly selected seven (10%) catchments were used for validation purposes and also seven for testing purposes. In case of BRANN, the validation catchments were not used, only testing have been done. Similarly to the MLR analysis, this too has also been used with different model setups (number of nodes, number of variables changing) to examine the robustness of the model.

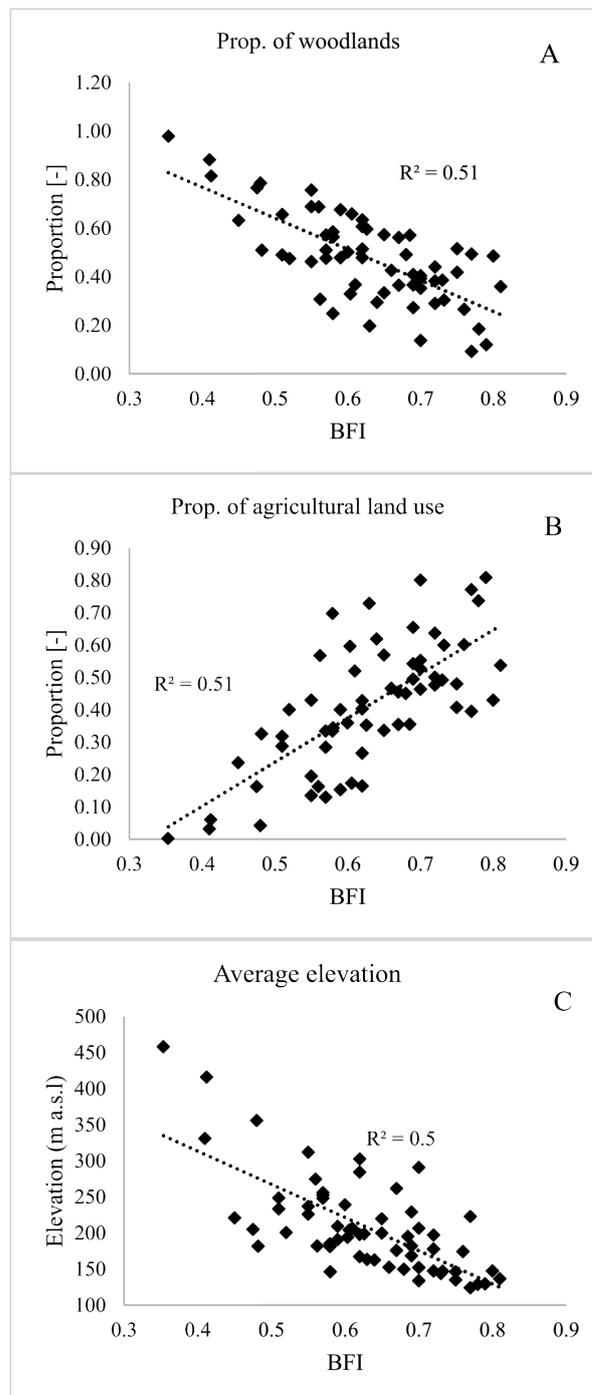
## 3 Results

### 3.1 Base flow indexes

BFI has been determined by base flow separation on 87 catchments over 100 and below 1000 km<sup>2</sup>, and additional 23 for smaller (below 100 km<sup>2</sup>) catchments with a minimum value of 0.35 and a maximum value of 0.92. The average BFI value is 0.67, which means that on average the two third of the total flow in (selected) rivers with entirely Hungarian catchments originate from subsurface runoff.

### 3.2 Simple regressions

Viewing the simple regressions between the variables and BFI, the best predicting variables seems to be the proportion of woodlands, proportion of agricultural land use, average elevation and sand content, while climatic variables also have evident connection with BFI values (Fig. 2, Table 2).



**Fig. 2** Best fitting simple regressions between catchment properties and base flow indices (a–c)

**Table 2** Pearson  $R^2$  values for catchment properties and base flow index

Catchment property	Pearson correlation ( $R^2$ )	Orientation
Woodlands prop.	0.51	(-)
Agricultural land. Prop.	0.51	(+)
Average elevation	0.5	(-)
Sand content	0.3	(+)
Avg. Temperature	0.25	(+)
Avg. yerly evapotr.	0.25	(-)
Solid rocks with poor por.	0.67	(-)
Catchment size	0.14	(+)

Except for catchment size, some kind of linear relationship between the variables and the BFI is evident. In case of the catchment area it seems that there are other factors that influence this relationship, therefore in its own it is not a clear correlation. The direction of correlation (Table 2) is also clear in most cases, proportion of woodlands, evapotranspiration, solid rocks with poor permeability and average elevation are clearly acting against recharge hence underground flows, while increasing sand content in the topsoil and the proportion of agriculture with much less interception capacity increase the recharge. Temperature is also positively correlated, which is probably because with lower elevation, temperature increases and slope decreases, hence the proportion of runoff will be smaller. Variables have been tested also with optimisation based variable selection.

### 3.3 Multiple Linear Regression analysis

The analysis gives different results, depending on which variables are kept in the group for analysis, and also on which catchments are used for training and validation.

The reason for this high variability of the results is the relatively low number of catchment used for the analysis, hence the varying catchment properties result in a different regression equation after the optimisation process. Forward and backward stepwise regressions resulted in poorer results, therefore best model search have been used for the prediction.

Rather than giving one single best equation for the regression (e.g. Fig. 3), the model was run 30 times, using 13 variables for input (without urban land use ratio) and used the best model search algorithm (which also changes the number of variables used). The resulting adjusted  $R^2$  was varying between 0.63 and 0.75, while the optimal number of variables was 7 on average (Table 3). The results are acceptable, with similar values to other studies [11,17]. The most influential variables (variables that were selected to the best model equations most times) were found to be: Catchment area, proportion of total agricultural area, proportion of solid rock areas with poor porosity, average elevation, average sand content, average annual evapotranspiration and average temperature.

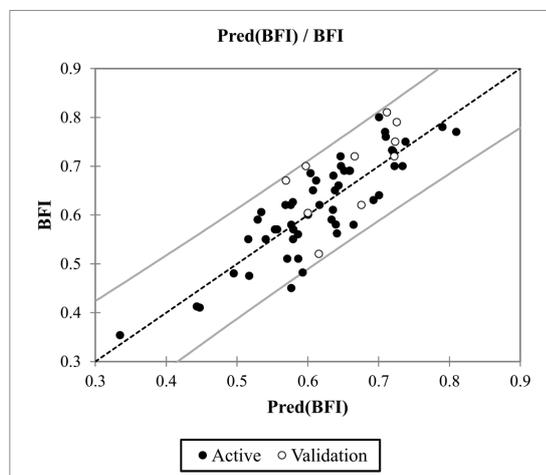


Fig. 3 One typical result of best fitting equation from 30 MLR analysis results including the results for validations and the 95% confidence intervals

Table 3 Error and Pearson correlation values for 30 best fitting model

No. of variables in the equation	Mean square error	$R^2$	Adjusted $R^2$
7	0.0038	0.72	0.68
8	0.0044	0.67	0.62
7	0.0035	0.72	0.68
6	0.0041	0.65	0.61
7	0.0037	0.69	0.65
8	0.0031	0.75	0.71
10	0.0028	0.80	0.75
7	0.0034	0.72	0.68
7	0.0033	0.72	0.67
6	0.0032	0.74	0.70
8	0.0029	0.74	0.70
7	0.0030	0.77	0.74
7	0.0033	0.72	0.68
8	0.0033	0.76	0.71
8	0.0033	0.69	0.63
6	0.0031	0.74	0.71
9	0.0035	0.75	0.70
6	0.0026	0.78	0.75
7	0.0032	0.76	0.72
9	0.0033	0.77	0.72
6	0.0030	0.74	0.70
6	0.0030	0.75	0.71
6	0.0031	0.74	0.71
6	0.0032	0.76	0.73
6	0.0026	0.77	0.74
6	0.0030	0.73	0.69
6	0.0030	0.73	0.70
8	0.0031	0.74	0.69
7	0.0027	0.79	0.76
7	0.0030	0.76	0.72

Six of these most influential variables have the best simple correlations with base flow index (Table 2), the only variable that did not get to the equation too often is the proportion of woodlands. Catchment area seems to be important despite the low simple linear correlation and that is because this is a fairly independent variable, with small cross correlation values with other properties. The likely reason why woodland is not in the best predictor variables, is that it has a strong cross correlation (-0.97) with average proportion of agricultural land use, therefore including both variables into the regression equation would not improve the results significantly.

Finding the best regression equation for further use is difficult with such a variability of equations with almost the same predictive power. Instead, it is proposed that the average of 30 model run results (best fitting equations) is taken as the best estimate with the indication of uncertainty (Fig. 5). Calculation by digital filter highlighted by strong contours, the rest is predicted by MLR. Uncertainties are indicated by dot

size. Before taking the average of the model runs, values that are outside the range of the mean  $\pm$  two times the standard deviation were removed. The result from this method is free of non-realistic values such as negative numbers, or numbers greater than 1. The goodness of fit can be also demonstrated by the mean square error, which is slightly above 0.003, and is reduced when the number of model runs, included in the average calculation, is increased (Fig. 4).

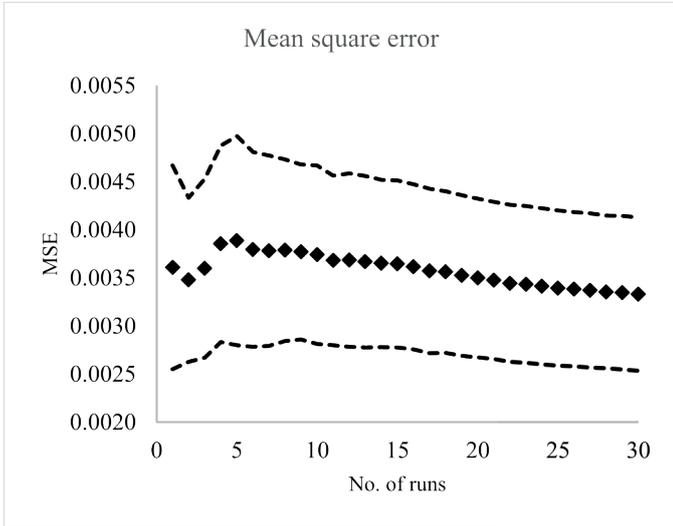


Fig. 4 Average of the Mean square error of the BFI estimation for the  $n^{\text{th}}$  number of runs showing the 95% confidence interval with dashed line

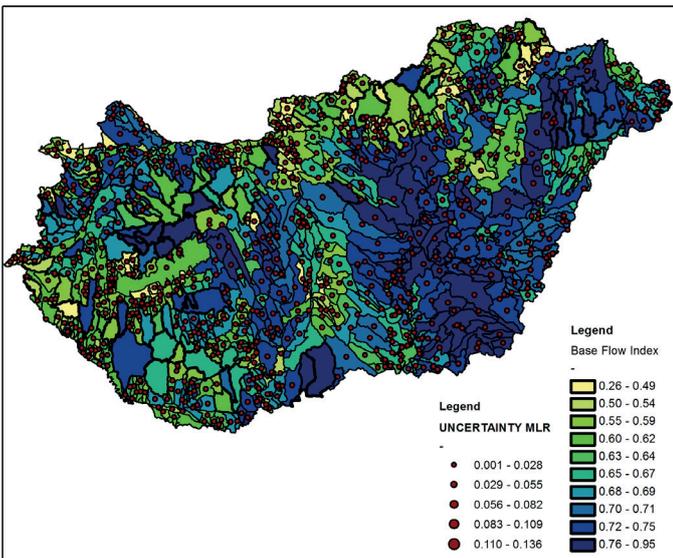


Fig. 5 MLR Predicted and calculated base flow index values for each catchment of Hungary

### 3.4 Artificial Neural Network analysis

During the testing of the networks and the optimisation methods, it was found that the LMA method was very sensitive to the number of nodes and the included variables in terms of the output, while the BRANN method proved to be robust. Therefore the latter has been selected for the prediction of the BFI-s. Neural network fitting to the training dataset can be considered acceptable with  $R^2$  values of 0.69 to 0.74 depending

on number of nodes and variables. This result is very much similar to the values of the linear regression results (Fig. 7) (Catchments with strong contours are from base flow separation results, the rest is by ANN prediction). The result is showing the prediction made using 3 hidden nodes and 13 variables (highest number examined), assuming that the highest number of variables can reflect better the natural variability as it contains more factors. The uncertainty concerning the mean of the model predictions with different setups is relatively small except a few areas, mostly large lakes and highly urban areas, which is sensible (Fig. 6). The BFI pattern for the country is quite similar to that of the MLR analysis as the difference map of the BFI values suggest (Fig. 7) (Beige colour shows good agreement, orange and red colour shows areas with higher prediction values for ANN, blue and grey for MLR).

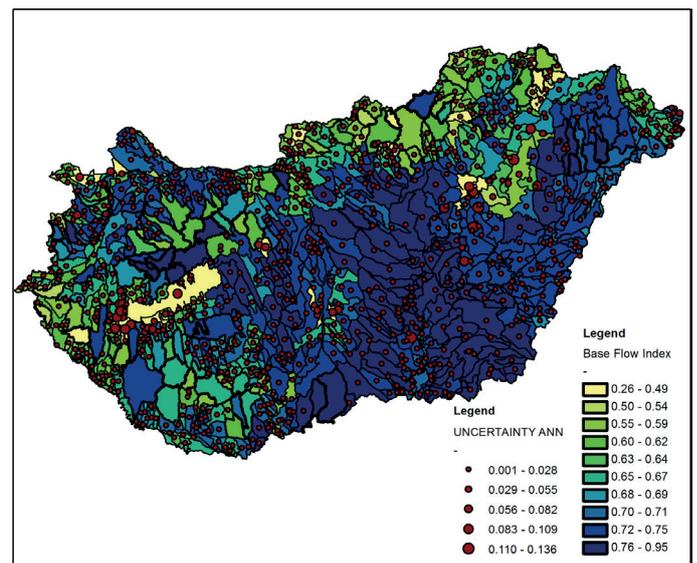


Fig. 6 ANN Predicted and calculated base flow index values for each catchment of Hungary

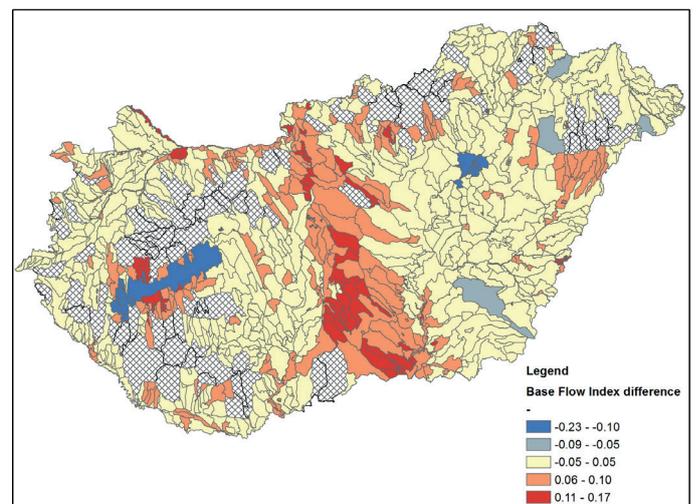


Fig. 7 Differences of BFI values calculated by the two methods (ANN-MLR)

The areas of highest concern with regard of the agreement between the two methods is the Danube-Tisza Interfluve and the hilly areas north of it. In mountainous areas the ANN analysis suggest values in the range of 0.5 to 0.7, which seems to be in agreement with the base flow separation values and in places (volcanic rocks at Börzsöny) it predicts values below 0.5 just as MLR analysis does. The mean square error of the BFI fitting is slightly better than that of the linear regression estimation, being mostly between 0.0025 and 0.003.

## 4 Discussion

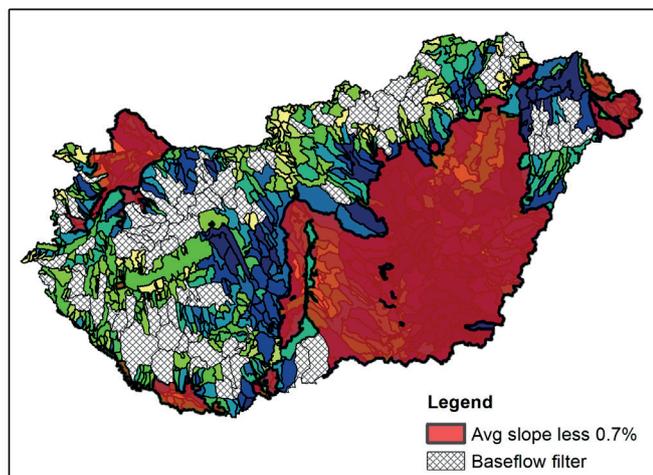
### 4.1 Base flow separation

The average value of BFI-s are quite far from the values calculated by the Moneris model [1], which has an average value of 0.63 for the whole country and 0.75 on average for the catchments used for base flow estimation. The greatest concern regarding the Moneris estimation method is however, that it estimates the surface runoff part of the water balance based on empirical formula, that is based only on the specific runoff from unpaved areas [1], while the ground water flow is given by the result after subtracting all the water flow components from total runoff, therefore input data bias has a large impact on the BFI results.

**Table 5** Average BFI values for gauged catchments by two methods

Slope class	Moneris	Base Flow sep.
Slope < 1%	0.78	0.64
1% <= Slope < 5%	0.72	0.68
5% <= Slope	0.73	0.79

It has to be mentioned that the number of gauged catchments with average slope less than 1% is very small, therefore this BFI value is not representative. Moreover, the smallest slope in the analysis is 0.71%, therefore catchments, where the average slope is less than that value (Fig. 8) are calculated by extrapolation, which might be still relevant if the linearities apply, however the uncertainty is higher at these catchments



**Fig. 8** Catchments with an average slope less than 0.7% and catchments where base flow filters have been used

### 4.2 Multiple linear regression

An interesting results of the regression analysis is that the inclusion of the proportion of the urban areas in the catchments as a predictor variable has improved the adjusted  $R^2$  values by 1–2 percent, and the variable has been selected into the regression equation by stepwise regression analysis more than half of all cases. One would expect that the proportion of urban lands would have a negative correlation with the base flow index [19] as runoff increases on impervious surfaces. In this analysis however on the contrary, the proportion of urban land use gets a positive coefficient in the MLR equation to predict the base flow index. This effect might reflect several different factors, such as garden irrigation, drinking water and sewer exfiltration, illegal or legal constant discharges. Despite this effect, urban land proportion has been left out from the regression analysis as it definitely does not work on catchments with large cities, as it would cause the Base Flow Index to rise above 1 for example in the case of Budapest (the regression might not work on fully urbanised catchments anyway).

The distribution of the base flow index tells that the low lying, flat areas of Hungary has high base flow contribution to the rivers, which is something that is expected. The highest portion of base flow in the total runoff is experienced on the karstic areas in the Trans-Danubian Mountains (Dunántúli-középhegység) in the mid-western part of the country and on the also karstic Northern Mountains (Bükk) based on base flow separation with values between 0.75 to 0.92. In the latter area however the regression suggest quite low values on some catchments (e.g. Szinva-patak), which might be unrealistic. The north-eastern sandy region (Nyírség) also show high portions of groundwater flow in the runoff (0.75–0.81), which is in a good agreement with the catchments assessed by base flow separation technique. Looking at two main basins in the Trans-Danubian region (Zala and Kapos) the BFI values range between 0.5 and 0.7 typically. The sand ridge between the Tisza and the Danube, the results are quite contradictory, as most of the region is covered by sand, therefore the variability shown on the map (Fig. 5) is not expected. This can be owed to the fact, that sand content is not an influential variable in the regression equations, only 13 times of the total 30 model runs does the best fitting equation contain sand content as a variable, but looking at separate runs, the picture is not very much different either. The other reason behind the relatively large variation in the base flows might be the difference in terrain topography and land cover. The regression predicts smaller base flow values in areas where the proportion of the woodlands are higher, which is quite sensible.

The lowest portion of groundwater in the river discharge is experienced in the volcanic mountains in the Northern-Mountains (Mátra, Börzsöny) with values below 50 % in many catchments, but in other catchments 60 percent could be marked as a typical value. The large woodland coverage and the poorly permeable volcanic rocks definitely explains low values.

The largest unknown as far as the BFI is concerned is the lowland areas of Hungary, namely the Great Plain (Alföld) in the Trans-Tisza region and the Small Plain area (Kisalföld) in the north-west. According to this analysis the BFI's are typically in the range of 0.7 and 0.85, while in places, where the soil condition and the land use is different e.g. at the Hortobágy the values are lower, around 0.6. This result however is considered uncertain as no catchments in this area has been estimated by base flow separation, due to lack of continuous flow measurements therefor this can be considered as an extrapolation and threatened as a preliminary value. The most important task in the future would be to provide measurements and validate, or contradict the current results.

Results for large cities, where the entire catchment is urban, should not be considered relevant, as many factors (i.e. illegal, legal point discharges, infiltration, storm runoff etc.) affecting base flow indexes have a different scale compared to rural catchments. Catchments where the proportion of lake area is higher than fifty percent, should also be considered as unreliable, as it is indicated by the uncertainty values (Fig. 5, Fig. 6).

#### 4.3 Artificial neural network

The ANN prediction suggest that the ratio of the base flow is generally large at the southern part of the Great-Plain and on the sandy ridge of the Danube-Tisza Interfluvium Fig. 6. The former is in agreement with the MLR results, while the latter is not entirely, the sandy regions show larger spatial variation for BFI by the MLR method and quite homogenous results by the ANN prediction with a mean value of around 0.78. The land cover of the mid-western to south-eastern parts of the interfluvium area is more heterogeneous with lots of forests, vineyards and pastures and less crop fields. It seems that the neural network relies more on the soil types, while the linear regression counts more on the land cover.

#### 4.4 Comparison of the two methods

The average BFI value for the predicted catchments are 0.63 and 0.66 for the MLR and ANN analysis respectively, while 0.77 for the MONERIS (without zero values). Standard deviation for MLR results is close to 0.1, the minimum BFI value being 0.26, the maximum being 0.86. The same values for ANN are 0.09, 0.36 and 0.8 respectively. MONERIS has a much wider range with 0.22 std.dev. and a range between zero and 0.9. The values suggest that ANN method calculates a narrower range for base flow indexes, while the regression simply extrapolate in some cases based on a linear equation.

### 5 Summary and conclusions

Country scale base flow map have been prepared for Hungary in small catchment scale to help calculations of water balance in country scale modelling of nutrient emissions on water bodies identified in the RBMP. The base flow index have been

calculated by digital filter based separation technique for all catchments that has at least one year long continuous river discharge data between 2001 and 2010. Based on the results, the BFI's have been regionalised on ungauged catchments using multiple linear regression and artificial neural networks.

Base flow indexes based on filter method show that some of the karstic areas, sandy places and also flat lands (only one gauged flat catchment) has the highest base flow portion in the river runoff in the range of 0.75 to 0.92. The hilly areas in the south-eastern countryside shows a larger variation between 0.5 and 0.75. The Northern Mountains shows a smaller variations around the mean value of 0.6, with only 2 exceptions towards high and low values.

The multiple linear regression analysis (best model selection, mean value of 30 runs) produced acceptable results with adjusted  $R^2$  values around 0.65–0.75, with 7 predictor variables on average. Six of the most frequent predictor variables are the ones that have the strongest separate correlation with BFI. Urban area proportion was not used as predictor despite the fact, that it is included in best equations many times, producing the highest adjusted  $R^2$  values. As the best equations varied as the validation catchments has been randomly selected, predictions have been averaged based on 30 best model selection procedure, also indicating the uncertainty of the predictions among the 30 runs.

The prediction showed comparable BFI values in areas around gauged catchments. In some cases, low values have been predicted on mountainous regions, with poorly permeable volcanic rock regions, which also seems reasonable. Some controversies have been found in karstic and sandy areas, which needs further investigation. The MLR predicted mostly high base flow values (0.7–0.8) on flat, low lying areas, but some difficult to explain variations have been also predicted, which needs justification. The same applies in the sand ridge region between the Danube and Tisza rivers.

ANN analysis show comparable results to the MLR analysis, its values being within 10% of the values of the MLR results in most of the country, but in the middle section of the country and in some other parts the two models show significant discrepancies. Uncertainty of the predictions are smaller than that of the MLR analysis, meaning that ANN is not so sensitive to the selection of the validation catchments or the predictor variables.

Looking at the variance of the two models, MLR results look more realistic, even if some of the results (on individual catchments) are unlikely. Comparison of the BFI values of the Moneris model and that of the digital filter separation on gauged catchments show differences on each major slope classes, with higher variation of the values based on the filter method. This suggest that the simple empirical formula that the Moneris model uses might be underestimating the natural variance of the subsurface/surface runoff ratio. This fact leads to the conclusion that the current BFI map might give more accurate results for runoff

components in general, even if in some specific catchments the BFI value might not be close to the reality due to the several factors that is not described well by the regression estimate.

Better validation of the current BFI estimation techniques is necessary and can be done through the implementation of continuous flow measurements on flat areas and through the applications of base flow separation measurements (tracer tests, multiple gages etc) on selected catchments on different geological regions.

## Acknowledgement

We thank the National Water Directorate of Hungary for providing discharge data of the national hydrological monitoring network and other data (DEM) that has been used for regression analysis.

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

## References

- [1] Venohr, M., Hirt, U., Hofmann, J., Opitz, D., Gericke, A., Wetzig, A., Natho, S., Neumann, F., Hürdler, J., Matranga, M., Mahnkopf, J., Gadegast, M. and Behrendt, H. "Modelling of Nutrient Emissions in River Systems - MONERIS - Methods and Background". *International Review of Hydrobiology*, 96(5), pp. 435–483. 2011. <https://doi.org/10.1002/iroh.201111331>
- [2] RIISAC. *AGROTOPO* (agronomical soil cover database). Budapest, 1991.
- [3] Keresztesi, Z., Kocsis, K., Schweitzer, F. "*Magyarország terkepeken: Domborzat es taj*". (*Hungary in maps: Topography and Landscape in Hungarian*). Budapest: MTA Földrajzi Kutatóintézet (Geographical Research Institute, Academy of Sciences), 2011.
- [4] Kozma, Zs. "*Belvízi szélsőségek kockázatalapú értékelésének és modellezési módszertanának fejlesztése*". (Development of the risk based evaluation and modeling methodology of excess water extremities. In Hungarian). Budapest: Vásárhelyi Pál Doctoral School, BUTE, 2014. [http://www.omikk.bme.hu/collections/phd/Epitomernoki\\_Kar/2014/Kozma\\_Zsolt/ertekezes.pdf](http://www.omikk.bme.hu/collections/phd/Epitomernoki_Kar/2014/Kozma_Zsolt/ertekezes.pdf)
- [5] *Országos Vízügyi Terv 2015* (National River Basin Management Plan 2015. In Hungarian), 2015. [http://www.vizugy.hu/vizstrategia/documents/E3E737A3-3EBC-4B6F-973C-5DD9B8A6DBAB/OVGT\\_foanyag\\_vegleges.pdf](http://www.vizugy.hu/vizstrategia/documents/E3E737A3-3EBC-4B6F-973C-5DD9B8A6DBAB/OVGT_foanyag_vegleges.pdf)
- [6] Jolánkai, Zs., Kardos, M., Muzelák B. "River Basin Management Plan of Hungary-annex 3.1". [www.vizeink.hu](http://www.vizugy.hu/vizstrategia/documents/988BF7DB-B869-46C6-9463-E9E4BFC81D2A/3_1_Hatteranyag_FEV_tapanyag_terhelemek_modellezes.pdf). 2015. [http://www.vizugy.hu/vizstrategia/documents/988BF7DB-B869-46C6-9463-E9E4BFC81D2A/3\\_1\\_Hatteranyag\\_FEV\\_tapanyag\\_terhelemek\\_modellezes.pdf](http://www.vizugy.hu/vizstrategia/documents/988BF7DB-B869-46C6-9463-E9E4BFC81D2A/3_1_Hatteranyag_FEV_tapanyag_terhelemek_modellezes.pdf)
- [7] EEA. *Corine Land Cover database*. <http://www.eea.europa.eu/data-and-maps/data/clc-2006-vector-data-version-2>, 2012
- [8] Kovács, Á. "*Tó és területi párolgás becslésének pontosítása és magyarországi alkalmazása*" (*Improving the estimation of lake and land evapotranspiration in Hungary*, in Hungarian). Budapest, Budapest University of Technology and Economics, 2009. <https://repozitorium.omikk.bme.hu/bitstream/handle/10890/1042/ertekezes.pdf?sequence=1&isAllowed=y>
- [9] Santhi, C., Allen, P. M., Muttiyah, R. S., Arnold, J. G., Tuppad, P. "Regional estimation of base flow for the conterminous United States by hydrologic landscape regions". *Journal of Hydrology*, 351(1–2), pp. 139–153. 2008. <https://doi.org/10.1016/j.jhydrol.2007.12.018>
- [10] Arnold, J. G., Allen, P. M. "Automated Methods For Estimating Baseflow and Ground Water Recharge From Streamflow Records". *JAWRA Journal of the American Water Resources Association*, 35(2), pp. 411–424. 1999. <https://doi.org/10.1111/j.1752-1688>
- [11] Ahiablame, L., Chaubey, I., Engel, B., Cherkauer, K., Merwade, V. "Estimation of annual baseflow at ungauged sites in Indiana USA". *Journal of Hydrology*, 476, pp. 13–27. 2013. <https://doi.org/10.1016/j.jhydrol.2012.10.002>
- [12] Partington, D., Brunner, P., Simmons, C. T., Werner, A. D., Therrien, R., Maier, H. R., Dandy, G. C. "Evaluation of outputs from automated baseflow separation methods against simulated baseflow from a physically based, surface water-groundwater flow model". *Journal of Hydrology*, 2012: 458–459. <https://doi.org/10.1016/j.jhydrol.2012.06.029>
- [13] Eckhart, K. "A comparison of baseflow indices, which were calculated with seven different baseflow separation methods". *Journal of Hydrology*, 352(1–2), pp. 168–173. 2008. [10.1016/j.jhydrol.2008.01.005](https://doi.org/10.1016/j.jhydrol.2008.01.005)
- [14] Nathan, R. J., McMahon, T. A. "Evaluation of Automated Techniques for Baseflow and Recession Analysis". *Water Resources Research*, 26(7), pp. 1465–1473. 1990. <https://doi.org/10.1029/WR026i07p01465>
- [15] Arnold, J. G., Allen, P. M., Muttiyah, R., Bernhardt, G. "Automated Baseflow Separation and Recession Analysis Techniques". *Groundwater*, 33(6), pp. 1010–1018. 1995. <https://doi.org/10.1111/j.1745-6584.1995.tb00046.x>
- [16] McCuen, R. H. "*Hydrologic Analysis and Design*". Prentice Hall, Upper Saddle River, New Jersey, 2004.
- [17] Mazvimavi, D., Meierink, A. M. J., Stein, A. "Prediction of base flows from basin characteristics: a case study from Zimbabwe". *Hydrological Sciences–Journal–des Sciences Hydrologiques*, 49(4), pp. 703–715. 2004. [http://hydrologie.org/hsj/494/hysj\\_49\\_04\\_0703.pdf](http://hydrologie.org/hsj/494/hysj_49_04_0703.pdf)
- [18] Rumsey, C. A., Millera, M. P., Susonga, D. D., Tillman, F. D., Anning, D. W. "Regional scale estimates of baseflow and factors influencing baseflow in the Upper Colorado River Basin". *Journal of Hydrology: Regional Studies*, 4(Part B), pp. 91–107. 2015. <https://doi.org/10.1016/j.ejrh.2015.04.008>
- [19] Bloomfield, J. P., Allen, D. J., Griffiths, K. J. "Examining geological controls on baseflow index (BFI) using regression analysis: An illustration from the Thames Basin, UK". *Journal of Hydrology*, 373(1–2), pp. 164–176. 2009. <https://doi.org/10.1016/j.jhydrol.2009.04.025>
- [20] Beck, H. E., van Dijk, A. I. J. M., Miralles, D. G., de Jeu, R. A. M., Bruijnzeel, L. A., McVicar, T. R., Schellekens, J. "Global patterns in base flow index and recession based on streamflow observations from 3394 catchments". *Water Resources Research*, 49(12), pp. 7843–7863. 2013. <https://doi.org/10.1002/2013WR013918>
- [21] Hocking, R. R. "The Analysis and Selection of Variables in Linear Regression". *Biometrics*, 32(1), pp. 1–49. 1976. <https://doi.org/10.2307/2529336>
- [22] Piotrowski, A. P., Napiorkowski, J. J. "A comparison of methods to avoid overfitting in neural networks training in the case of catchment runoff modelling". *Journal of Hydrology*, 476, pp. 97–111. 2013. <https://doi.org/10.1016/j.jhydrol.2012.10.019>
- [23] Chow, V. "*Handbook of Applied Hydrology*". New York, USA: McGraw-Hill, 1964.
- [24] Werbos, P. J. "*Beyond regression: New tools for prediction and analysis in the behavioral sciences*". Cambridge, MA, USA: Ph.D. Thesis, Harvard University, 1974.
- [25] Madsen, K., Nielsen, H. B., Tingleff, O. "Methods for Non-Linear Least Squares Problems". *Technical University of Denmark*, 2004. [http://www2.imm.dtu.dk/pubdb/views/edoc\\_download.php/3215/pdf/imm3215.pdf](http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/3215/pdf/imm3215.pdf)
- [26] Burden, F., Winkler, D. "Bayesian regularization of neural networks". In: *Artificial Neural Networks. Methods in Molecular Biology*. (Livingstone, D. J. (Ed.)). pp. 25–44. 2008. [https://doi.org/10.1007/978-1-60327-101-1\\_3](https://doi.org/10.1007/978-1-60327-101-1_3)