2014

Rediscovering Manning'S Equation Using Genetic Programming

Carlos F. Gaitan

11th International Conference on Hydroinformatics
HIC 2014, New York City, USA

REDISCOVERING MANNING'S EQUATION USING GENETIC PROGRAMMING

CARLOS F GAITAN (1)
(1): South Central Climate Science Center, University of Oklahoma, 201 Forrestal Road, Princeton NJ 08540 USA

Open-channel hydraulics' research traditionally links empirical formulas to observational data, for example Manning's formula for open channel flow (Q) driven by gravity relates the cross-sectional average velocity (V), the hydraulic radius (R), and the slope of the water surface (S) with a friction coefficient n, characteristic of the channel's surface. Here we use novel Genetic Programming (GP), a technique inspired by nature's evolutionary rules, to derive empirical relationships based on synthetic datasets of the aforementioned parameters. Specifically, we evaluated if Manning's formula could be retrieved from datasets with 300 pentads of A, n, R, S, and Q (from Manning's equation). The cross-validated results show success retrieving the functional form from the synthetic data (even in the presence of an uncorrelated predictor) and encourage the application of GP on problems where traditional empirical relationships show high biases or are non-parsimonious. The results also show alternative flow equations that can be used in the absence of one or more predictors and that approximate Manning's equation.

INTRODUCTION

With growing data complexity and an increasingly high amount of observations and model simulations within the geosciences, the discovery of new scientifically significant relationships could be daunting given the dimensions of these big-datasets [1]. However, techniques from other disciplines like computer science, economics and bioinformatics can often be used to tackle common problems in hydrological sciences. In particular, novel fields like climate informatics and hydro-informatics relate climate and hydrological sciences, respectively, with approaches from statistics, machine learning and data mining. These disciplines, inspired by the advances in computer science and bioinformatics during the last 30 years, can provide innovative ways of analyzing data and of extracting knowledge from data collections.

There are numerous studies using artificial intelligence/machine learning methods to solve problems in hydrology, climatology and geosciences. For example, Ghosh and Mujumdar [2] downscaled stream-flow using relevance vector machines, Toprak and Cigizoglu [3] predicted longitudinal dispersion coefficients in natural streams using different types of neural networks,

Coulibaly, Dibike [4] forecasted non-stationary hydrological time series using dynamically driven recurrent neural networks, Francke, López-Tarazón [5] used quantile regression forests to determine sediment transport, Zeng, Hsieh [6] used support vector regression to predict seasonal winter extreme precipitation over Canada, Gaitán, Hsieh [7] compared linear and nonlinear regression models when downscaling maximum and minimum temperatures, and Guistolisi [8] used genetic programming to determine the Chezy coefficients in corrugated channels. Similarly, Tang, Reed [9] tested different multi-objective evolutionary algorithms for hydrologic model calibration, and showed that a strength Pareto evolutionary algorithm attained competitive results when used to calibrate the Sacramento soil moisture accounting model for the Leaf River watershed, and when calibrating an integrated hydrological model for the Shale Hills watershed in Pennsylvania (USA). However, Babovic and Abbot's [10, 11] two-part document (The evolution of equations from hydraulic data) constitutes the first antecedent of the use of evolutionary algorithms for hydraulic modeling, sediment transport, salt water intrusion in estuaries, and in flow resistance studies.

Similarly, open-channel hydraulics' (OCH) research often links empirical formulas to observational data (e.g. Weisbach (1845), St. Venant (1851), Neville (1860), Darcy and Bazin (1865)). For example, the Manning formula, also known as the Gauckler-Manning-Strickler formula (hereafter GMS), is an empirical formula for open-channel flow, or free surface flow driven by gravity. The formula is attributed to the engineers Philippe Gauckler (1967), Robert Manning (1890) and Albert Strickler (1923). The formula (1) relates the cross-sectional average velocity (V=Q/A), the hydraulic radius (R), and the slope of the water surface (S), with a friction coefficient n, characteristic of the channel's surface.

$$V = (1/n)\, R^{2/3} S^{0.5} \quad (1)$$

Where, V is the cross-sectional average velocity in m/s, n is a non-dimensional roughness coefficient, R is the hydraulic radius (m), and S is the slope of the water surface (m/m). The relationship can be used to calculate the discharge (Q) if we substitute V in (1) by Q/A, obtaining:

$$Q = (A/n)\, R^{2/3} S^{0.5} \quad (2)$$

Research involving the GMS equation traditionally focuses on the determination of the roughness coefficient under different flow regimes (e.g. Ayvaz [12] and Ding, Jia [13]) and/or for different riverbed materials (e.g. Candela, Noto [14]), as even the presence of biological soil crusts can affect the surface roughness, runoff and erodibility of the channel [15].

Our goal is to retrieve the GMS equation from synthetic hydraulic data, and to evaluate alternative solutions with varying degrees of complexity using novel genetic programming (GP). As with Darwin's induction method, where the hypothesis comes from analyzing the data, genetic programming generates possible solutions that fit the data given an evaluation metric. The adaptation of these solutions to the data is akin to the biological adaptation of an individual member of a population to an environment. The solutions' equations are obtained by randomly combining different building blocks (operators). These operators are typically algebraic (+, –, ÷, ×), trigonometric (e.g. sin(x), cos(x), tanh(x)), or conditional (e.g. if statements). However, other functions typically used in computer programs can also be used [16]. In general, GP abandons unviable solutions (offspring) and retains viable ones. The

solutions are usually evaluated in terms of fitness functions such as mean absolute error (MAE), correlation coefficient, and Bayesian Information Criterion (BIC), among many others; and the algorithm stops when a desired accuracy level is reached.

METHODS & DATA

Genetic Programming is an evolutionary computation technique that automatically solves problems without requiring the user to know or to specify the form of the solution in advance [19]. As stated by Poli, Langdon [16], GP is a systematic, domain independent method for getting computers to solve problems automatically. Similarly, if one considers Darwin's adaptation theory as the accumulation of knowledge about an environment [10], GP's solutions represent adapted solutions to the data. In general terms, GP uses evolutionary operators like crossover and mutation. Crossover creates two offspring solutions by combining randomly chosen parts from two selected parent solutions, while mutation creates a child/offspring solution by randomly altering a randomly chosen part of the selected parent solution [16].

To create the programs, the user determines a priori function sets and terminal sets that could be part of the final solution (offspring); examples of function sets include arithmetic, mathematic, boolean, and conditional functions, among many others. On the other hand, a terminal set from which all end (leaf) nodes in the parse trees representing the programs must be drawn. Examples of terminal sets include variables, constants and functions without arguments [20].

Here we used 300 instances of four different predictors (A, R, S and n) and the corresponding 300 values of Q (calculated using equation 2). To generate more parsimonious solutions with the GP tool, we opted to use the following building blocks: constant, addition, subtraction, multiplication, division and power. Hence avoiding trigonometric functions like sine and cosine, often used when a periodic signal is expected (e.g. seasonal cycle). To obtain the possible solutions we used Eureqa™ 0.99.4 Beta [21] and kept its default values for the initial population size, stopping criteria and cross-validation characteristics. We archived non-optimal solutions to aid the evolving programs to discover common intermediate states and converge to them, following the recommendation of Krawiec [22]. The software algorithm also controls the maturity and the stability of the proposed solutions. Where maturity measures how long ago the top solutions last improved, and stability measures how long it has been since any solution improved.

The model complexity is computed by summing the number of times a particular type of expression (i.e. variable, real number, +, -) appears in an expression weighted by the building block complexity (e.g. 1 for constants, multiplications and additions; 2 for divisions; 3 for sines and cosines, 4 for tangents; and 5 for power operations).

Data

Our experimental setup includes two experiments. The first one uses synthetic variables of A, R, S and n, with the corresponding Q - from the GMS equation - using the data intervals shown in Table 1. While the second experiment uses the data ranges in Table 2, and an uncorrelated variable generated using seasonal cycle anomalies of 2 m temperature from the 64X13Y

NCEP/NCAR reanalysis gridpoint [23]. The NCEP/NCAR dataset was obtained through Environment Canada's DAI portal [24].

Variable	Range	Step size
A	1 – 3.98	0.01
R	0.25-30.05	0.1
S	0.00025-0.03005	0.0001
N	0.009-0.07456	0.00022

Table 1. Predictor variables used in experiment one.

With the first experiment we wanted to show if the new GP-generated equations created overfitted solutions that worked only on a small subsample of the data, as we used a group of data points with Q values below 4 m^3s^{-1} for training, and tested the models with data points outside this interval; we also wanted to know if the GP tool was able to obtain the exact functional form of the GMS equation. With the second experiment we tested GP's ability of to select relevant predictors.

Variable	Range	Step size
A	1 – 443.54	0.01+noise
R	0.25-175.57	0.1+noise
S	0.00025-0.03005	0.0001+noise
n	0.009-0.07456	0.00022+noise

Table 2. Predictor variables used in experiment two.

Results

The following results correspond to the best-performing models obtained by the GP environment, as the evolutionary process described in the introduction involves the creation of a large number of (potential) expressions, involving multiple offspring and generations (iterations). In particular, for the first experiment, we trained the GP models on a subset of data points with Q < 4 m^3s^{-1} and tested the models with a subset of points outside that interval. The GP-generated equations in Table 3 include the top 6 top solutions that worked well within the 0-4 m^3s^{-1} range, as the proposed solutions have low mean absolute errors (MAEs) and high (~1) correlation coefficient. However, when evaluating the models performance outside of the aforementioned range, only the first two models were general enough to work outside the training interval. Models 3, 4, 5, 6 likely represent overfitted solutions and should only be used when the predictors are inside of the training interval.

Model Number	Model solution
1	$A\,(R^{0.990})^{0.673}/nS^{-0.5}$
2	$AR^{0.667}/(nS^{-0.5})$
3	$1.11An^{-0.949}S^{0.502}$
4	$A/n\,(1.12S)^{-0.511}$
5	$A^{0.558}(73.9S)^{0.595}$
6	$12(AS)^{0.582}$

Table 3. Results using the variables from Table 1.

Overall, two models (5 and 6) used A and S as predictors, two models (3 and 4) used A, S and n as predictors, and the other two models (1 and 2) used A, S, n and R as predictors. Numerically,

the solution of model 2 represents the GMS solution, while model 1 is a less parsimonious version of it.

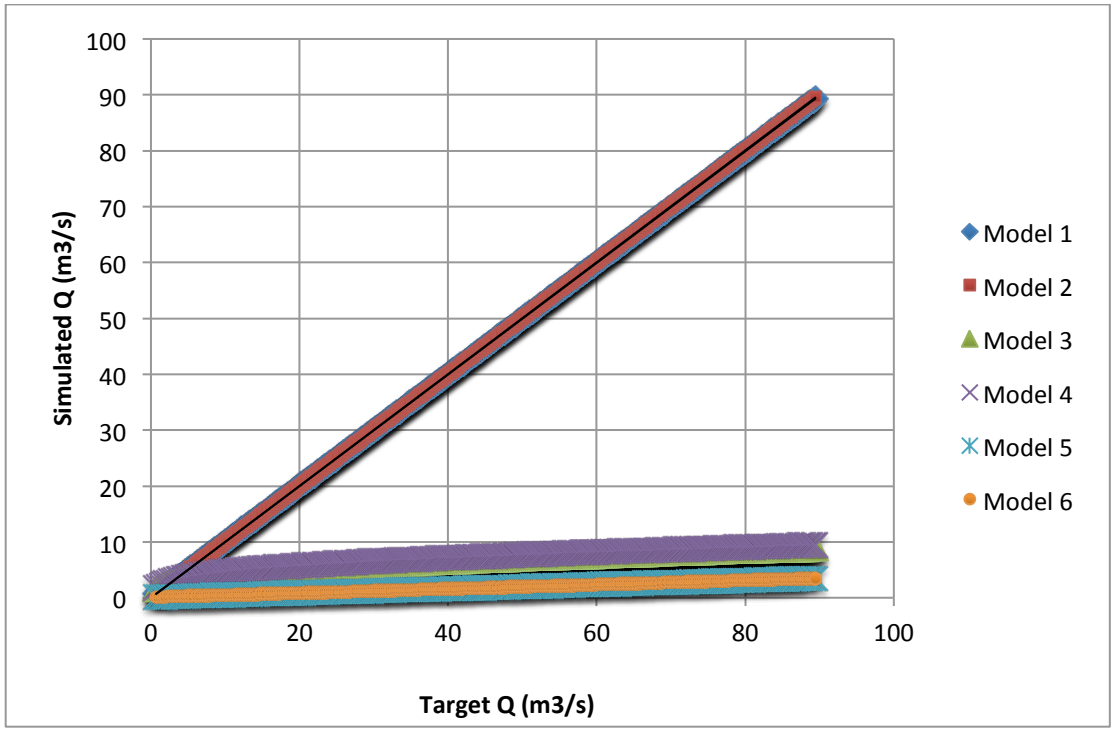

Figure 1 Results using the variable ranges shown in Table 2.

Now that we obtained different formulas that approximate the GMS equation, it is important to balance the equation taking into consideration that both sides of the proposed solutions should have the same units (i.e. m^3s^{-1}). For example, the right hand side of M1's equation has to be multiplied by a factor k = 1 $m^{1/3}/s$, so the equation has flow units. Table 4 shows the equations' coefficients and their units.

ID	Model solution	Coefficient	Units
GMS	$1\ A\ R^{0.667}S^{1.5}/(n\ S)$	1	$m^{1/3}s^{-1}$
M1	$A\ R^{0.667}S^{1.5}/(n\ S)$	1	$m^{1/3}s^{-1}$
M2	$(0.0082\ A^{1.487}R^{0.6812})/n$	0.0082	$m^{-0.4843}s^{-1}$
M3	$(0.01\ R\ A^{1.184})/n$	0.01	$m^{-0.2721}s^{-1}$
M4	$(0.005368\ A^{2.134})/n$	0.0058	$m^{-1.3893}s^{-1}$
M5	$10.58\ A^{1.319}$	10.58	$m^{0.5051}s^{-1}$
M6	$2.602e^6S^2/n$	2.602	m^3s^{-1}

Table 4. Proposed solutions, coefficients and their SI units.

The Pareto chart in figure 2 shows in the y-axis the MAE between the target and the simulation and in the x-axis the model complexity –as explained in the Data and Methods section -. In this figure, the top solutions can be found at the bottom right corner, where lower MAEs from less complex models are located. The results show that only two solutions obtained no errors, the GMS and the M1 models. The difference in complexity between the GMS and the M1 solutions is caused by the absence of the 1 $m^{1/3}/s$ coefficient in M1, unlike the GMS solution.

Overall, our results show that the GP technique was able to retrieve the original form of the GMS equation when using different synthetic datasets; additionally the GP methodology proposed new, alternative solutions that can approximate the GMS original equation (for values of Q less than 40000 m^3s^{-1}). These new solutions are often more parsimonious than the GMS equation and require fewer parameters, but their MAEs versus the GMS solution were between 27 and 589 m^3s^{-1}. Finally we included in Table 4 the selected GP-generated equations and the units of their corresponding coefficients in order to have dimensionally balanced equations. The constants found in these solutions usually have dimensions of m^Xs^{-1}, with X varying between -1.39 (for M4) and 3 (for M6), where the GMS solution includes X equal 1/3.

4. Discussion and Recommendations

Here we show novel equations for OCH generated using genetic programming. The new proposed equations can impact immediately OCH's research and offer parsimonious approximations of the GMS equation for free surface flow driven by gravity. Additionally, we showed a new application of GP in hydrological sciences and corroborated the ability of GP methods of retrieving the functional form of the equation that generated the data.

We used genetic programming and implemented two genetic programming operations: mutation and crossover, to detect nonlinear equations of open channel hydraulics, in various synthetic datasets derived from the GMS equation. The analytical solutions that we found included the original relationship, together with more parsimonious and more complex solutions, often involving a fewer number of predictors. However, even thou the method suggested promising expressions that approximated the GMS equation, it also suggested over-fitted expressions that worked only in certain intervals, as seen in figure 1.

As mentioned by Graham, Djorgovski [1], automated discovery methods, like genetic programming can be applied to any general dataset, and many potential applications can be found in fields where theoretical gaps exist despite abundance in data [17], as this kind of techniques may help the scientists to focus on other interesting phenomena more rapidly and to interpret their meaning. This characteristic is especially appealing when dealing with big-datasets, like the ones found in hydrology, climatology, astronomy and other geophysical sciences.

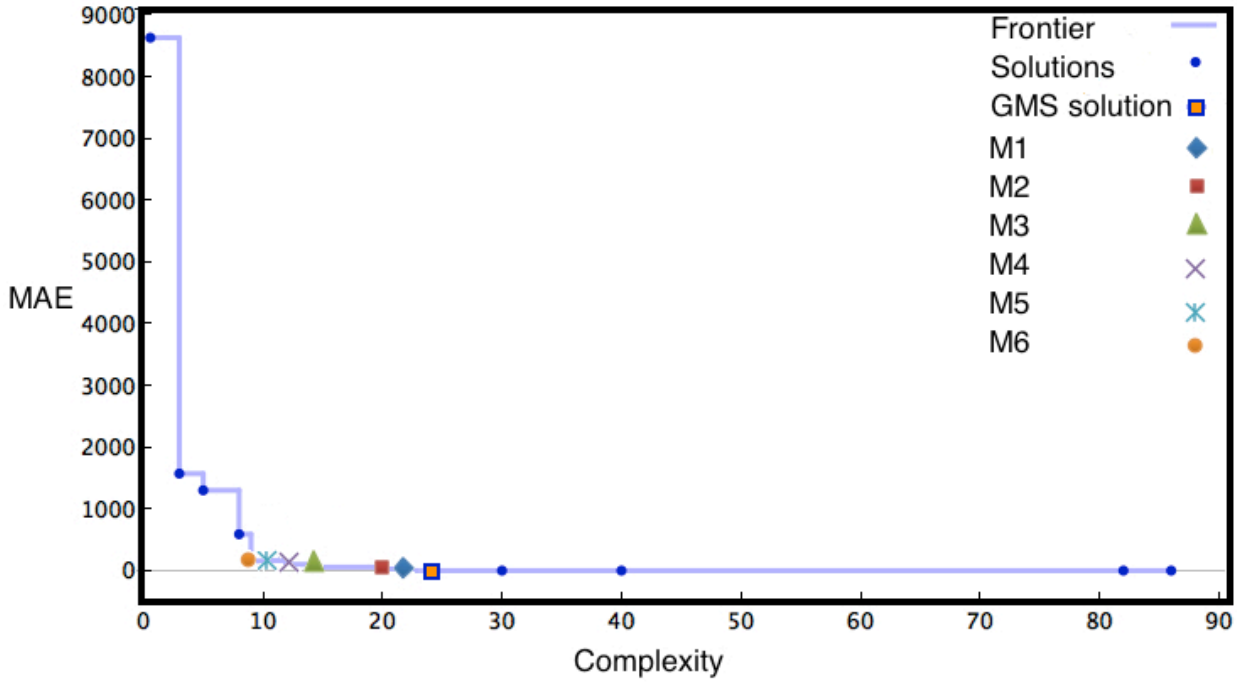

Figure 2. Pareto chart of the different GP-generated solutions. Dark dots indicate possible solutions. The solutions marked with different identifiers correspond to the best equations, including the GMS solution (square).

On the other hand, we found that the proposed GP technique could be used for feature selection and extraction, as the method successfully omitted unrelated variables –like 2m temperature- from the proposed equations. This suggests that the method could also be used for nonlinear predictor selection, complementing classical methods like the stepwise selection, often used in conjunction with multiple linear regression, and provides an alternative to the graphical sensitivity analysis method by Cannon and McKendry [25] and to the Bayesian approach used by Robertson and Wang [26] for seasonal streamflow forecasting.

Finally, we did a dimensional analysis on all the GP-generated models, including the ones omitting some of the GMS predictors so these alternative solutions that can be used in the absence of certain explanatory variables or when the data quality of the predictors is compromised -as observations errors can heavily impact the output of hydrological and hydraulic studies [27]-. Future applications include (but are not limited to): a) predictor selection in statistical downscaling, b) determination of empirical relationships between river flow and suspended sediments, c) calibration of soil moisture functions, d) generation of alternative evapotranspiration equations, and d) creation of alternatives to the empirical equations that determine the watershed time of concentration (i.e the time required for the runoff to travel from the hydraulically most distant point to the outlet).

Acknowledgements

The authors would like to acknowledge the Data Access Integration (DAI) Team for providing the data and technical support. The DAI Portal (http://loki.qc.ec.gc.ca/DAI/) is made possible through collaboration among the Global Environmental and Climate Change Centre (GEC3),

the Adaptation and Impacts Research Division (AIRD) of Environment Canada, and the Drought Research Initiative (DRI).

REFERENCES

1. Graham, M.J., et al., *Machine-assisted discovery relationships in astronomy.* MNRAS, 2013. **431**(3): p. 2371-2384.
2. Ghosh, S. and P.P. Mujumdar, *Statistical downscaling of GCM simulations to streamflow using relevance vector machine.* Advances in Water Resources, 2008. **31**(1): p. 132-146.
3. Toprak, Z.F. and H.K. Cigizoglu, *Predicting longitudinal dispersion coefficient in natural streams by artificial intelligence methods.* Hydrological Processes, 2008. **22**(20): p. 4106-4129.
4. Coulibaly, P., Y.B. Dibike, and F. Anctil, *Downscaling precipitation and temperature with temporal neural networks.* Journal of Hydrometeorology, 2005. **6**(4): p. 483-496.
5. Francke, T., et al., *Flood-based analysis of high-magnitude sediment transport using a non-parametric method.* Earth Surface Processes and Landforms, 2008. **33**(13): p. 2064-2077.
6. Zeng, Z., et al., *Surface Wind Speed Prediction in the Canadian Arctic using Nonlinear Machine Learning Methods.* Atmosphere-Ocean, 2011. **49**(1): p. 10.
7. Gaitán, C.F., et al., *Evaluation of Linear and Non-Linear Downscaling Methods in Terms of Daily Variability and Climate Indices: Surface Temperature in Southern Ontario and Quebec, Canada.* Atmosphere-Ocean, 2013: p. 1-11.
8. Guistolisi, O., *Using genetic programming to determine Chezy resistance coefficient in corugated channels.* Journal of Hydroinformatics, 2004. **0.6**(3): p. 157-173.
9. Tang, Y., P. Reed, and T. Wagener, *How effective and efficient are multiobjective evolutionary algorithms at hydrologic model calibration?* Hydrol.Earth Syst.Sci, 2006. **10**: p. 289-307.
10. Babovic, V. and M.B. Abbott, *The evolution of equations from hydraulic data Part I: Theory.* Journal of Hydraulic Research, 1997. **35**(3): p. 397-410.
11. Babovic, V. and M.B. Abbott, *The evolution of equations from hydraulic data Part II: Applications.* Journal of Hydraulic Research, 1997. **35**(3): p. 411-430.
12. Ayvaz, M.T., *A linked simulation–optimization model for simultaneously estimating the Manning's surface roughness values and their parameter structures in shallow water flows.* Journal of Hydrology, 2013. **500**: p. 183-199.
13. Ding, Y., Y. Jia, and S.S.Y. Wang, *Identification of Manning's Roughness Coefficients in Shallow Water Flows.* Journal of Hydraulic Engineering, 2004. **130**: p. 501-510.
14. Candela, A., L.V. Noto, and G. Aronica, *Influence of surface roughness in hydrological response of semiarid catchments.* Journal of Hydrology, 2005. **313**(3-4): p. 119-131.
15. Rodríguez-Caballero, E., et al., *Effects of biological soil crusts on surface roughness and implications for runoff and erosion.* Geomorphology, 2012. **145-146**: p. 81-89.
16. Poli, R., W.B. Langdon, and N.F. McPhee, *A Field Guide to Genetic Programming*. 2008, UK.
17. Schmidt, M. and H. Lipson, *Distilling Free-Form Natural Laws from Experimental Data.* Science, 2009. **324**: p. 81-85.
18. Pappa, G.L., et al., *Contrasting meta-learning and hyper-heuristic research: the role of evolutionary algorithms.* Genetic Programming and Evolvable Machines, 2013.
19. Koza, J.R., *On the programming of computers by means of natural selection*. 1996: MIT Press.
20. Langdon, W.B., *Genetic Programming and Data Structures*. 1996, London: University College.
21. Schmidt, M., *Eureqa User Guide*. 2011.
22. Krawiec, K., *Genetic Programming: where meaning emerges from program code.* Genetic Programming and Evolvable Machines, 2013.
23. Kistler, R., et al., *The NCEP-NCAR 50-year reanalysis: Monthly means CD-ROM and documentation.* Bulletin of the American Meteorological Society, 2001. **82**(2): p. 247-267.
24. DAI_Team, *Catalogue of Available Datasets Through DAI.* 2008, Environment Canada. p. 25.
25. Cannon, A.J. and I.G. McKendry, *A graphical sensitivity analysis for statistical climate models: application to Indian monsoon rainfall prediction by artificial neural networks and multiple linear regression models.* International Journal of Climatology, 2002. **22**(13): p. 1687-1708.
26. Robertson, D.E. and Q.J. Wang, *A Bayesian approach to predictor selection for seasonal streamflow forecasting.* Journal of Hydrometeorology, 2012. **13**(1): p. 155-171.
27. Di Baldassarre, G. and A. Montanari, *Uncertainty in river discharge observations: a qualitative analysis.* Hydrol.Earth Syst.Sci, 2009. **13**: p. 913-921.