8th IAHR
EUROPE
CONGRESS
LISBON - PORTUGAL
4-7 JUNE 2024

Addressing class imbalance problems in data-driven rainfall-runoff modelling

Federico VILASECA[1], Christian CHRETIES[1], Alberto CASTRO[2,3] and Angela GORGOGLIONE[1]

[1] Institute of Fluid Mechanics and Environmental Engineering (IMFIA), Universidad de la República, Uruguay
[2] Institute of Computer Science (INCO), Universidad de la República, Uruguay
[3] Institute of Electrical Engineering (IIE), Universidad de la República, Uruguay
fvilaseca@fing.edu.uy; chreties@fing.edu.uy; acastro@fing.edu.uy; agorgoglione@fing.edu.uy

ABSTRACT

This paper proposes a methodology based on data augmentation to improve the performance of data-driven rainfall-runoff models on high flows. Problems in the representation of such range of discharges by data-driven models were presented in previous research, which the authors of this work attribute to class imbalance of discharge data, where high flows are underrepresented. This ends up biasing the learning process towards the representation of low flows. The proposed methodology was tested for two incremental watersheds of the Santa Lucía Chico basin in Uruguay, showing an increase in performance of 17 % for Nash-Sutcliffe efficiency and 38 % for peak-flow Nash-Sutcliffe efficiency. Results demonstrate that class imbalance is a relevant issue affecting the performance of data-driven rainfall-runoff models and that the proposed methodology allows to tackle said issue and improve model performance for high flows.

1. Introduction

Data-driven rainfall-runoff models serve as valuable tools in hydrological studies, offering insights into the complex relationships between rainfall patterns and runoff dynamics. However, recent investigations have highlighted their limited representation of flood peaks in daily step implementations (Barbosa-Reis et al., 2021; Chen et al. 2023; Rezaie-Balf et al., 2019, Vilaseca et al., 2023). We theorize that these limitations can often be attributed to class-imbalance issues present in both input and output time series. Typical hydrographs encountered in alluvial rivers show prevailing low and mid-flow conditions, occasionally punctuated by rapid streamflow increase generating flood events. Classifying flow events based on their magnitude places floods in the minority category, starkly contrasting with the abundant base flows. Consequently, this class imbalance distorts the learning process of algorithms during the training phase, impairing the model's ability to accurately predict and capture high-flow events. To address this issue, in this work, we present a methodology to balance the weight of high flows, with respect to base flows, during the training stage of data-driven rainfall-runoff models. By ameliorating the class imbalance, we aim to enhance the model's capacity to effectively predict and characterize flood events, contributing to the refinement of hydrological simulations and bolstering our understanding of the intricate interplay between precipitation and runoff dynamics.

2. Materials and methods

2.1. Study area and data

The study area is the Santa Lucía Chico basin, located in the south-central region of Uruguay. The entire watershed (2478 km^2) with closure at Paso Severino (PS) dam and an incremental watershed (1748 km^2) with closure at the city of Florida (FL) were considered to test the methodology. Available data included daily time series of river discharge, accumulated rainfall and air temperature, covering the period 1989-2016. These data were used to build two input datasets (A and B) for the data-driven models (Table 1). Further details about the study basin together with data description and analysis can be found in Vilaseca et al. (2023).

2.2. Workflow description

A series of 88 experiments were conducted per watershed, each of which consisted of the training and evaluation of a Random Forest model. For each, the input dataset was built out of the two possible variable combinations (Table 1), and randomly split into training and testing subsets with a 75/25 ratio. Then, a data augmentation workflow was applied to the training set to increase the weight of the high-flow data. Each experiment was carried out considering different: (1) input dataset, (2) method for classification of events, (3) data augmentation algorithm, (4) increase ratio of the amount of high flow (minority class) events.

Table 1. Input datasets for the models. MAP = mean areal precipitation, $\text{MAP}_{\text{accum}}$ = 7-day accumulated MAP, Tmax = max. temperature, Tmin = minimum temperature, Q = discharge, Q_{t-I} = Q of the i^{th} previous day

Dataset	Input variables	Output variable
Dataset A	MAP, $\text{MAP}_{\text{accum}}$, Tmax, Tmin	Q
Dataset B	MAP, $\text{MAP}_{\text{accum}}$, Tmax, Tmin, Q_{t-1}, Q_{t-2}	Q

Events were classified with two possible alternatives: peaks-over-threshold (POT) or k-means. The POT classification consisted of identifying local maximums in the discharge series over a fixed threshold of 500 m^3/s. This led to a binary event classification where flood peaks represented the minority class. The alternative was using the k-means clustering algorithm for unsupervised classification of the events. In this case, values of k = 2, 3, 4, or 5 were considered. Possible data augmentation algorithms included random sampling with replacement from the known high flow (minority class) events or generation of new synthetic minority class events using four variants of the SMOTE (Chawla et al., 2002) algorithm: SMOTE, Borderline SMOTE (Han et al., 2005), SVM SMOTE (Nguyen et al., 2009) or ADASYN (He et al., 2008). Each augmentation algorithm was set to increase the amount of minority class events in a ratio of $IR = N^*/N$, being N the number of events of such class prior to the data augmentation and N^* said number after the augmentation. IR was set to take values between IR = 1.5, 2.5, 5.

During the training process for each model, a selection of the Random Forest hyperparameters was optimized using four-fold cross-validation and the genetic algorithm Optuna (Akiba et al., 2019), following the same pipeline described in Vilaseca et al. (2023). The objective function for optimization was Nash-Sutcliffe efficiency (NSE), and models were evaluated for percent bias (PBIAS), ratio of mean squared error to standard deviation (RSR), NSE calculated for peaks (pkNSE), and NSE calculated for log-transformed flows (logNSE).

3. Results and discussion

The results of the best experiment compared to the baseline model (without data augmentation in the training set) for each watershed and each of the two input datasets are shown in Table 2.

Table 2. Results of the best experiments compared to baseline models for each watershed and dataset.

Watershed	Input dataset	Experiment	k (k-means)	Sampling method	IR	NSE	PBIAS	RSR	logNSE	pkNSE
PS	Dataset A	Baseline	-	None	-	0.34	0.17	1.32	-0.06	0.22
PS	Dataset A	Best	4	SMOTE	2.5	0.46	-17.6	0.94	-0.07	0.41
PS	Dataset B	Baseline	-	None	-	0.62	3.01	0.81	0.76	0.54
PS	Dataset B	Best	4	SMOTE	1.5	0.76	-2.41	0.6	0.66	0.71
FL	Dataset A	Baseline	-	None	-	0.38	4.69	1.42	0.25	0.29
FL	Dataset A	Best	4	SMOTE	2.5	0.39	-11.99	1.12	0.18	0.34
FL	Dataset B	Baseline	-	None	-	0.64	-3.87	0.73	0.85	0.58
FL	Dataset B	Best	3	SVM SMOTE	1.5	0.71	-10.77	0.64	0.85	0.66

Results show an improvement in the estimation of the flow peaks after applying the proposed data augmentation methodology. This is indicated by the increase in NSE and pkNSE indicators. It should also be noted that performance decreases for low flows, as evidenced by the comparison of logNSE indicators. In most cases, the best results are obtained for the SMOTE algorithm, with the k-means algorithm set for k = 4 clusters.

4. Conclusions

Results allow to conclude that the posed hypothesis is accurate, and that class imbalance is a relevant issue for data-driven rainfall-runoff modeling. The proposed method allows improving the representation of high flows based on known data augmentation techniques, while slightly lowering the performance for low flows, which shows a balancing of the learning process. On average, performance increased by 17% for NSE and 38% for pkNSE and decreased by 14% for logNSE.

Acknowledgments

This work was made possible thanks to a Ph.D. scholarship granted to F.V. by the Academic Postgraduate Commission (CAP), Universidad de la República, Uruguay.

References

Akiba T, Sano S, Yanase T, Ohta T, Koyama M (2019) Optuna: a next-generation hyperparameter optimization framework, KDD' 19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2623-2631.

Barbosa G, da Silva DD, Fernandes EI, Moreira MC, Vieira G, de Souza M, Rocha SA (2021) Effect of environmental covariable selection in the hydrological modeling using machine learning models to predict daily streamflow, Journal of Environmental Management, 290, 112625.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, 16, 321-357.

Chen Z, Lin H, Shen G (2023) TreeLSTM: A spatiotemporal machine learning model for rainfall-runoff estimation, Journal of Hydrology: Regional Studies, 48, 101474.

Han H, Wang W, Mao B (2005) Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning, ICIC 2005: Proceedings of the International Conference on Intelligent Computing, 878-887.

He H, Bai Y, Garcia BE, Li S (2008) ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning, Proceedings of the International Joint Conference on Neural Networks (IJCNN 2008), 1322-1328.

Nguyen H, Cooper EW, Kamei K (2009) Borderline over-sampling for imbalanced data classification, Proceedings of the Fifth International Workshop on Computational Intelligence and Applications, 24-29.

Rezaie-Balf M, Nowbandegani SF, Samadi SZ, Fallag H, Alaghmand S (2019) An ensemble decomposition-based artificial intelligence approach for daily streamflow prediction, Water, 11, 709.

Vilaseca F, Chreties C, Castro A, Gorgoglione A (2023) Assessing influential rainfall-runoff variables to simulate daily streamflow using Random Forests, Hydrological Sciences Journal, DOI: 10.1080/02626667.2023.2232356.