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

ABSTRACT

This paper proposes a methodology based on data augmentation to improve the performance of data-driven hydrological models during high flows. Problems in the representation of high discharges by data-driven models were observed in previous research, which the authors of this work attribute, in part, to the shortage of high-flow observations in the training data. This creates an imbalance problem that biases the learning process towards the representation of low flows. The proposed methodology was tested for models generated with the Random Forest machine learning algorithm, implemented in two incremental watersheds of the Santa Lucía Chico basin in Uruguay. Results showed an average increase in performance of 18 % for Nash-Sutcliffe efficiency and 37 % for peak-flow Nash-Sutcliffe efficiency. The work allows us to conclude that class imbalance is a relevant issue affecting the performance of data-driven rainfall-runoff models under certain conditions and that the proposed methodology is useful to tackle it, potentially improving model performance for high flows.

1. Introduction

Data-driven models serve as valuable tools in hydrological studies, offering insights into the relationship between rainfall patterns and river discharges. 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; Matos et al., 2018). Although high discharges can inherently be more difficult to model than low ones in absolute terms due to the common heteroscedastic behavior of river flows, we theorize that the issue of limited representation of flood peaks can often be attributed to class-imbalance issues in the time series employed for model training, at least in part. Typical hydrographs encountered in alluvial rivers show prevailing low and mid-flow conditions, occasionally punctuated by rapid streamflow increase generating flood events. Therefore, 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 capture and predict high-flow events accurately.

Diverse data augmentation techniques have been developed and applied through the years to data science-related problems in different areas of knowledge. They serve a variety of purposes but are particularly useful for the task of balancing datasets with unequal representation of classes, such as the problem posed in this paper. Among frequently used data augmentation methods, the Synthetic Minority Oversampling Technique (SMOTE) algorithm and its variants stand out for their ability to generate synthetic data close to the one existing in the multi-dimensional space of the input variables. To the authors' knowledge, data-augmentation techniques have been scarcely applied to hydrological modeling, and the class-imbalance problem in the discharge series has not yet been fully addressed. Some examples of data augmentation applications in hydrological models are the works of Bi et al. (2020) and Zhang and Yan (2023), who used linear interpolation to artificially increase the sampling resolution of discharge observations, resulting in a more robust training dataset and, consequently, improved model performance. Recent applications to other related water resources problems include the work of Tang et al. (2022) who combined SMOTE with machine learning-based models for precipitation forecasting, and of Bilali et al. (2021) who employed Gaussian noise to build synthetic

datasets in the implementation of a model for predicting fecal coliforms in rivers using the AdaBoost machine learning algorithm.

To address the class-imbalance issue in developing a machine learning-based hydrological model, we present a methodology based on the SMOTE algorithm and its variants to balance the weight of high flows, with respect to base flows, during the training stage in the implementation. The methodology is then applied to the development of Random Forest (RF) models for two incremental watersheds in the Santa Lucía Chico basin, Uruguay. 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.

2. Materials

2.1. Study area

The study area is the Santa Lucía Chico basin, located in the south-central region of Uruguay. It is characterized as a mainly rural watershed with mild terrain slopes. Its predominant land use is agriculture, with its territory majorly covered by pastures (82.4%) and crops (9.4%) (Vilaseca et al., 2021a). Located in a zone of temperate climate, the annual rainfall ranges between 1000 and 1500 m, with a regime characterized by high intra-annual variability, with no distinguishable rainy season. The air temperature varies during the year between 3 and 30 °C with a four-season regime.

The motivation for selecting this basin as the case study is its national relevance. Besides its significance in terms of agricultural production, it is the country's primary source of raw water for potabilization. It contains the Paso Severino reservoir (at the outlet), an artificial lake with a design volume of 70 hm^3, which stores water for the system that supplies drinking water to the capital city, Montevideo, and its metropolitan area. It serves more than half of Uruguay's population. An incremental watershed (1748 km^2), with outlet at the city of Florida (*FL*), and the entire watershed (2478 km^2), with outlet at Paso Severino (*PS*) dam, were considered to test the methodology (Figure 1). In both cases, outlets correspond to locations of discharge monitoring sites.

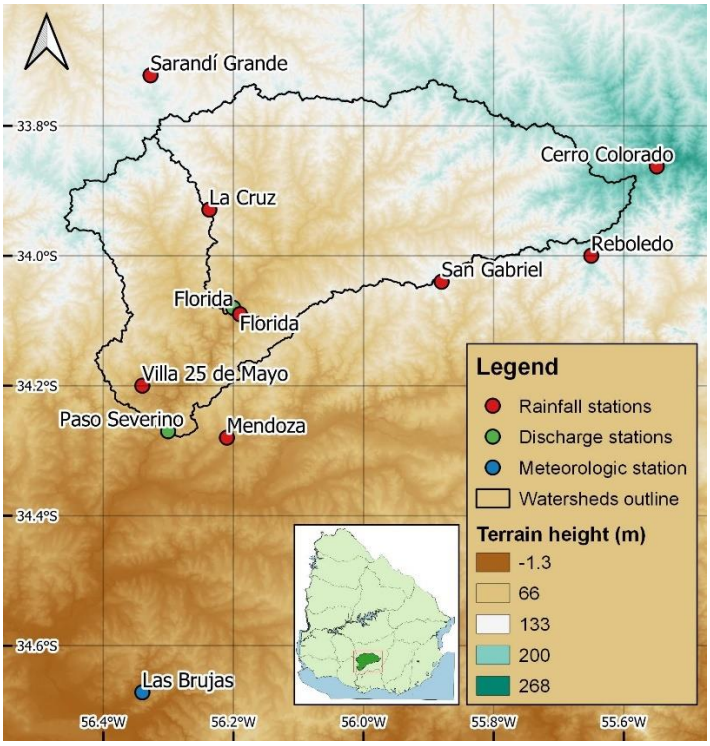

Figure 1. Location of the incremental watersheds under study and data stations around them.

2.2. Data

The dataset employed to build the models included time series of daily accumulated rainfall (*P*), in *mm*, measured at eight stations located around the basin; also maximum and minimum daily air temperatures (*Tmax* and *Tmin*), in *°C*, at one climatologic station situated in close proximity to the basin; and mean daily discharges (*Q*), in m^3/s, measured at the closure points of both incremental watersheds (FL and PS). In both cases, discharges are estimated from real-time water level measurements. In FL, through a rating curve, while in PS,

through theoretical level-discharge relationships both for the dam's spillway and the intake pipes that are operated to allow the passage of water through said structure (both outputs spillway and pipes are added to obtain the average daily discharge). The shared period of all the time series was from 1/17/1989 to 5/6/2016 (9973 observations).

3. Methods

3.1. Data preprocessing

Two sets of data were generated to be used as input for implementing the models. They were formed by combining the time series of original variables described in section *2.2* with other synthetic variables, introduced to facilitate the learning process and improve model performance. A time series of spatially averaged rainfall (*MAP*) was generated for each watershed using the Thiessen polygons method. This weighted rainfall series replaced the eight original ones to reduce the dimensionality of the dataset, which increases model performance while reducing overfitting (Vilaseca et al., 2023). In addition, a series of 7-day accumulated rainfall (MAP_{accum}) was calculated from the *MAP* series to be introduced as a *proxy* of the antecedent moisture state of the watershed. Lastly, time-lagged discharge time series of 1 (Q_{t-1}) and 2 days (Q_{t-2}) were also introduced as model inputs. The two sets of data that were built included spatially averaged rainfall, its 7-day accumulation and maximum and minimum temperatures, while one of them also included the two series of lagged discharges, as shown in Table 1.

Table 1. Input datasets for the models.

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

3.2. Random Forest models

RF (Breiman, 2001) is a widely used tree-based machine learning algorithm. It is an ensemble learning method since it consists of a set of decision trees individually trained with different subsets of the training dataset, which are generated through bootstrapping. When predicting, the output of the RF model is the mean of the outputs of each tree. This ensemble scheme has been proven to alleviate the bias issues presented by decision trees trained from the entire dataset. Due to their simplicity, quickness of implementation, performance, robustness, inbuilt uncertainty estimation, and interpretability, RF models are commonly applied to hydrological modeling, and that is the main reason to include them in the present study. Implementation was carried out in python using the scikit-learn library (Pedregosa et al., 2011)

3.3. Classification of discharge observations

Events were classified with two possible alternatives: peaks-over-threshold (POT) or k-means. The POT classification consisted of the identification local maxima in the discharge series over a fixed threshold with a value equivalent to the 99th percentile of the total time series. This led to a binary event classification where flood peaks represented the minority class. The alternative was using the k-means clustering algorithm (Forgy, 1965; Lloyd, 1982) for unsupervised classification of the events. The algorithm requires a number "k" of desired clusters to be set beforehand. It starts by generating a set of "k" centroids, randomly located in the multi-dimensional space determined by the variables of the dataset. Then, it optimizes the location of those centroids so that they match, in the best possible way, the distribution of the observations which are later classified according to their nearest centroid. In this case, values of k = 2, 3, 4, or 5 were considered, and all the variables (input and output) were given to the algorithm as input. High or peak discharges were associated with the class with fewer observations in it (minority class). Both classifying methodologies were implemented in python, using the scikit-learn library (Pedregosa et al, 2011) for the k-means algorithm.

3.4. Data augmentation algorithms

Possible data augmentation algorithms included random sampling with replacement (RRS) from the known high flow events (minority class) or generation of new synthetic minority class events using four variants of the SMOTE algorithm: SMOTE (Chawla et al., 2002), Borderline SMOTE (Han et al., 2005), SVM SMOTE (Nguyen et al., 2009) or ADASYN (He et al., 2008). The acronym SMOTE stands for Synthetic Minority Over-Sampling Technique. It is, as its name suggests, an algorithm that generates new synthetic observations inside the multi-dimensional space determined by the original dataset. In our case, it is applied only to high-

flow observations (minority class) to increase the volume of data in that range. The algorithm works by randomly choosing an observation and an associated number of nearest neighbors (3, in this case). Then, one of those neighbors is chosen, and the synthetic observation is created at a random point of the line determined by both selected observations. This procedure is repeated until the target amount of synthetic data is generated.

SMOTE variants are modifications of the original algorithm, in which the observations to oversample are not selected randomly. Instead, certain zones of the multi-dimensional space are prioritized. Borderline SMOTE prioritizes observations located close to the classification border, which means those whose classification has uncertainty. SVM SMOTE is an improvement of Borderline SMOTE in which the nearest neighbors of the chosen observation are identified through the SVM (Support Vector Machine) algorithm, seeking nearby observations that are also part of the classification border. Lastly, ADASYN (Adaptive Synthetic Sampling) gives preference to the zones of the multi-dimensional space with less density of observations.

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 of IR = 1.5, 2.5, 5. All the augmentation methods were performed using the imbalanced-learn python library (Lemaitre et al., 2017), except for RRS, which was implemented using scikit-learn (Pedregosa et al., 2011).

3.5. Full workflow description

A series of 150 experiments were conducted per watershed, each of which consisted of the implementation of an RF model through the following steps. A graphical scheme is shown in Figure 2 for better understanding.

1) Define the dataset to use from one of the two input variable combinations presented in Table 1.
2) Classify each observation using one of the alternatives described in *3.3*: POT or k-means, with *k* taking values of 2, 3, 4, or 5.
3) After the classification, randomly split it into training and testing subsets with a 75/25 ratio.
4) Perform the augmentation procedure for the training subset, using one of the algorithms detailed in *3.4* (RRS, SMOTE, Borderline SMOTE, SVM SMOTE, or ADASYN) to artificially increase the number of observations of the minority class (high flows), using a predefined IR chosen between 1.5, 2.5 or 5.
5) Train the RF model using the augmented training dataset. During the training process, a selection of the Random Forest hyperparameters is optimized for Nash-Sutcliffe efficiency (NSE) using four-fold cross-validation and the tree-structured Parzen estimator algorithm within the optimization framework Optuna (Akiba et al., 2019), following the same procedure described in Vilaseca et al. (2023).
6) Evaluate the model by making predictions for the testing period and comparing them to the observations in the testing dataset. Performance indicators are percent bias (PBIAS), ratio of mean squared error to standard deviation (RSR), NSE calculated for peak discharges (pkNSE), and NSE calculated for log-transformed flows (logNSE).

After the experiments, results were compared to discuss which alternatives are better for improving model performance without a loss of generalization capacity due to overfitting issues. To set a baseline for comparison, base models were also implemented for each watershed and dataset (4 in total). They were trained using the corresponding dataset in its original form without classifying observations or augmenting the data.

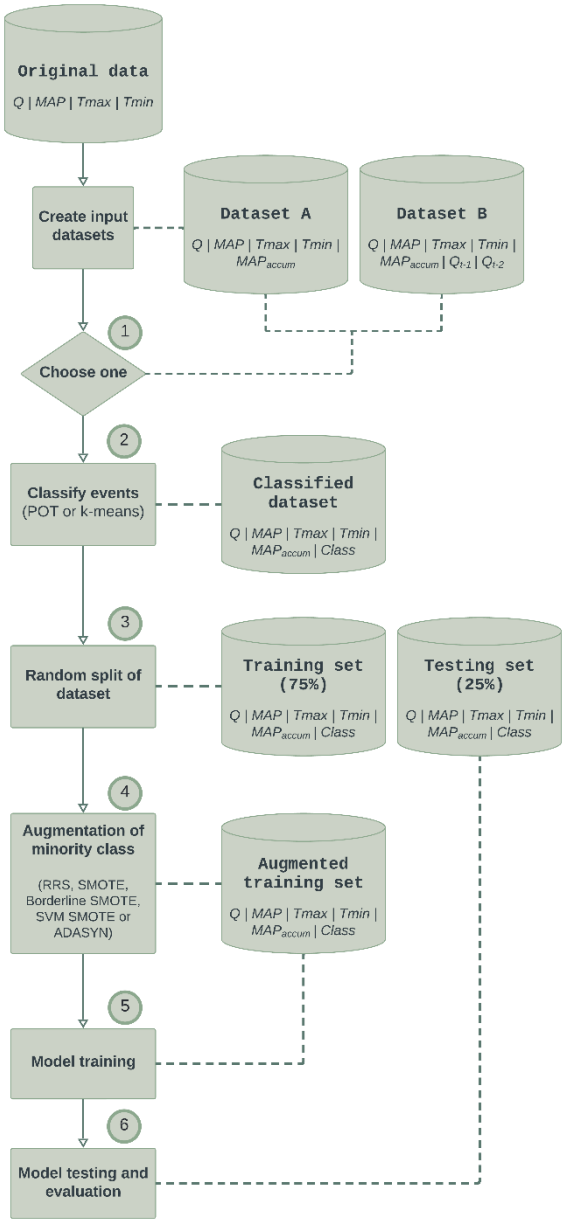

Figure 2. Flow diagram of the procedure for each experiment.

4. Results

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. Also, scatterplots of modeled *vs*. observed discharges comparing baseline to best models for each of the watershed and input dataset pairs are displayed in Figures 3-6.

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

Watershed	Input dataset	Experiment	Classification method	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	k-means	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	k-means	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	k-means	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	k-means	3	SVM SMOTE	1.5	0.71	-10.77	0.64	0.85	0.66

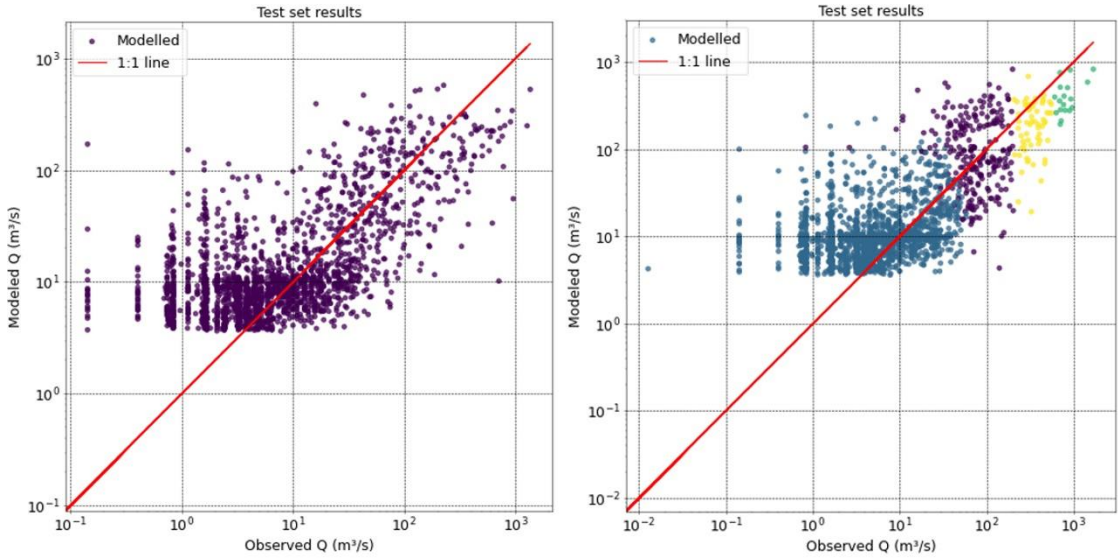

Figure 3. Comparative results of baseline (left) and best (right) models for Dataset A in Paso Severino watershed. The colours in the right image correspond to the classes assigned to each observation, which in this case was obtained with the k-means method (4 clusters).

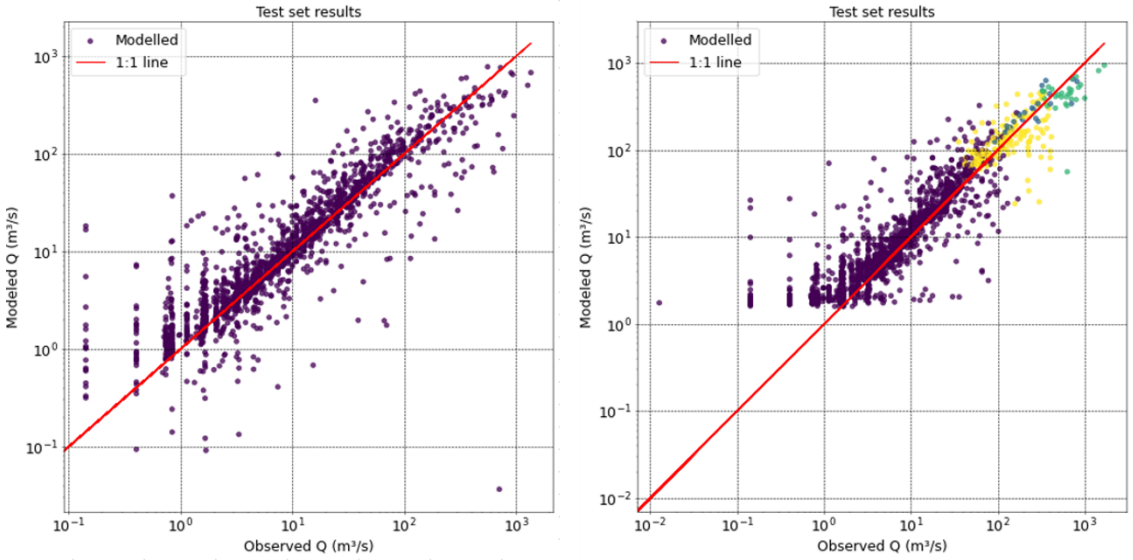

Figure 4. Comparative results of baseline (left) and best (right) models for Dataset B in Paso Severino watershed. The colors in the right image correspond to the classes assigned to each observation, which in this case was obtained with the k-means method (4 clusters).

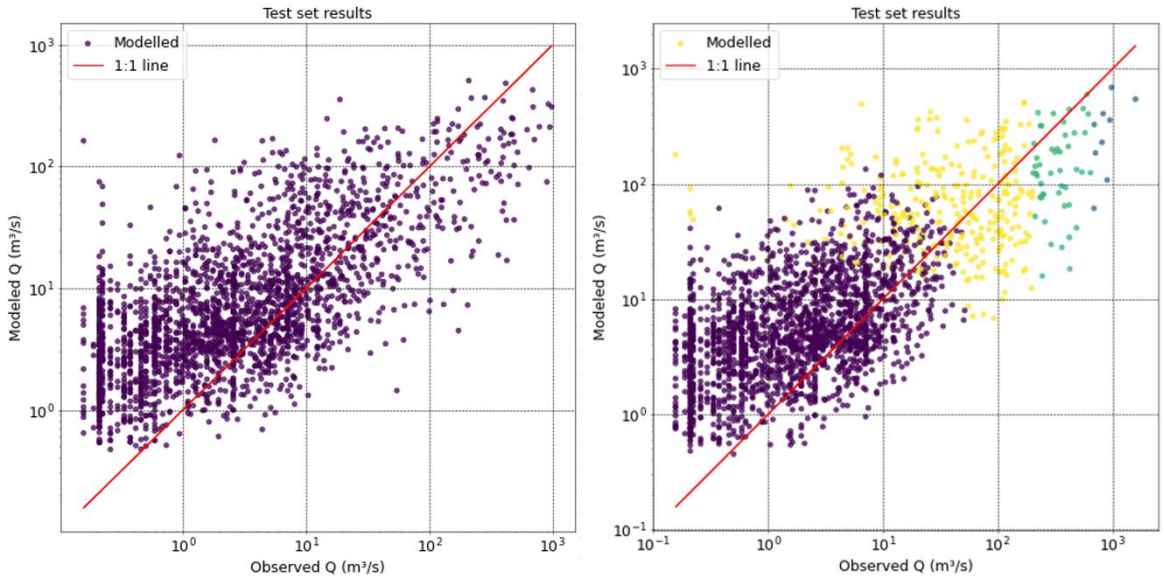

Figure 5. Comparative results of baseline (left) and best (right) models for Dataset A in Florida watershed. The colors in the right image correspond to the classes assigned to each observation, which in this case was obtained with the k-means method (4 clusters).

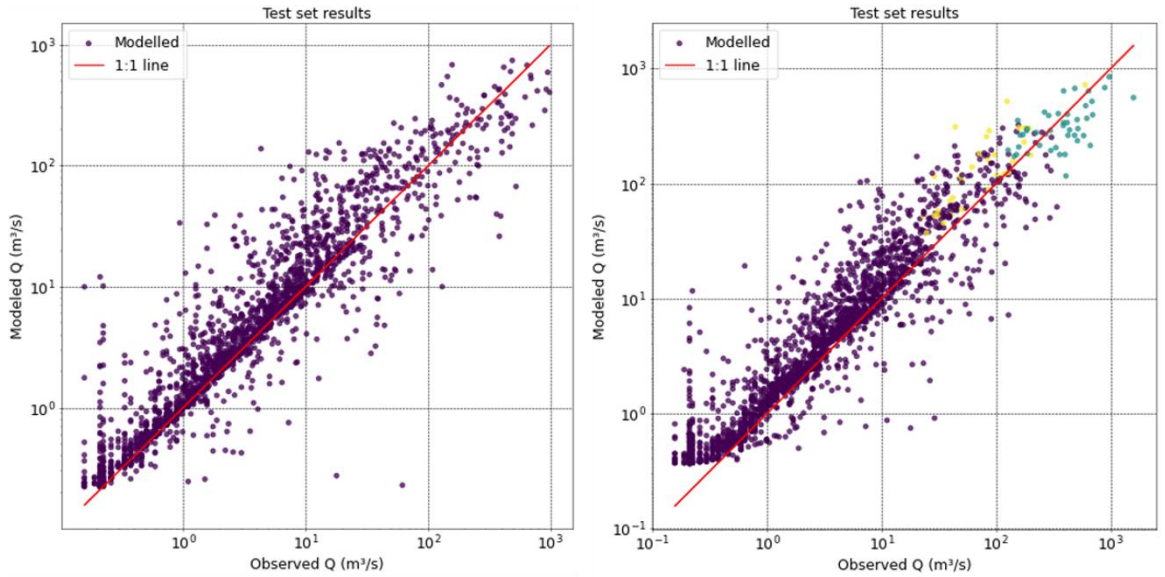

Figure 6. Comparative results of baseline (left) and best (right) models for Dataset B in Florida watershed. The colors in the right image correspond to the classes assigned to each observation, which in this case was obtained with the k-means method (3 clusters).

Both statistical and graphical results show an improvement in estimating high flows after applying the proposed data augmentation methodology, which is more notorious in PS than in FL. 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. Both observations can also be appreciated in the scatterplot comparison. In most cases, the best results are obtained for the SMOTE algorithm, with the k-means algorithm set for k = 4 clusters.

5. Discussion

Table 1 shows a significant improvement in model performance for high flows when applying the data augmentation methodology to the training dataset. This is reflected in the NSE and pkNSE indicators, which increase for the four combinations of watersheds and datasets. While NSE is typically used as an indicator of overall performance for hydrological models, it has been shown to be considerably affected by high flows (Clark et al., 2021), so it is also a valid indicator of increased performance for such range. NSE increases between 3% and 35% (mean 18%), while pkNSE does it from 14% to 86% (mean 37%). On the other hand, logNSE, which is a well-known indicator of performance for low discharge values, decreases in 3 out of the 4 cases, with percentages between -13% and -28% (mean -14%), and presents no significant change in the remaining one.

The scatterplot comparison allows conclusions corroborating the analysis of the indicators of Table 1. It can be seen in all cases (but more strongly in Paso Severino) that the point cloud, in the range of high discharges, converges more closely to the 1:1 line than in the plots of the right (best model) than in the ones of the left (baseline model). In the same way, they diverge from the 1:1 line when looking at the low-flow range.

6. Conclusions

Results allow concluding that the posed hypothesis is accurate, and that class imbalance is a relevant issue for data-driven rainfall-runoff modeling. The proposed method improves the representation of high flows based on known data augmentation techniques while slightly lowering the performance for low flows, which balances the learning process. On average, performance increased by 18% for NSE and 37% 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

Bi XY, Li B, Lu WL, Zhou XZ (2020) Daily runoff forecasting based on data-augmented neural network model, Journal of Hydroinformatics, 22(4), 900-915

Bilali AE, Taleb A, Bahlaoui MA, Brouziyne Y (2021) An integrated approach based on Gaussian noises-based data augmentation method and AdaBoost model to predict faecal coliforms in rivers with small dataset, Journal of Hydrology, 599, 126510

Breiman L (2001) Random forests, Machine learning, 45, 5-32.

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

Clark MP, Vogel RM, Lamontagne JR, Mizukami N, Knoben WJ, Tang G, Gharari S, Freer JE, Whitfield PH, Shook KR, Papalexiou SM (2021) The abuse of popular performance metrics in hydrologic modeling, Water Resources Research, 57(9), e2020WR029001

Forgy EW (1965) Cluster analysis of multivariate data: efficiency versus interpretability of classifications, Biometrics, 21, 768-769

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

Lemaitre G, Nogueira F, Aridas CK (2017) Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning, Journal of Machine Learning Research, 18(17), 1-5

Lloyd S (1982) Least squares quantization in PCM, IEEE transactions on information theory, 28(2), 129-137

Matos JP, Portela MM, Schleiss AJ (2018) Towards safer data-driven forecasting of extreme streamflows: an example using support vector regression, Water Resources Management, 32, 701-720.

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

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python, Journal of Machine Learning Research, 12, 2825-2830

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

Tang T, Jiao D, Chen T, Gui G. (2022) Medium-and long-term precipitation forecasting method based on data augmentation and machine learning algorithms, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, 1000-1011

Vilaseca F, Castro A, Chreties C, Gorgoglione A (2021) Daily rainfall-runoff modeling at watershed scale: a comparison between physically-based and data-driven models, In International Conference on Computational Science and Its Applications, 18-33

Vilaseca F, Chreties C, Castro A, Gorgoglione A (2023) Assessing influential rainfall-runoff variables to simulate daily streamflow using Random Forests, Hydrological Sciences Journal, 68(12), 1738-1753
Zhang J, Yan H (2023) A long short-term components neural network model with data augmentation for daily runoff forecasting, Journal of Hydrology, 617, 128853