EGU General Assembly 2020

Comparison of Machine Learning Techniques Powering Flood Early Warning Systems

Application to a catchment located in the Tropical Andes of Ecuador

Paul Muñoz[1,*] , Johanna Orellana-Alvear[1,2] , Jörg Bendix[2] and Rolando Célleri[1,3]

[*] **paul.munozp@ucuenca.edu.ec**
[1]Department of Water Resources and Environmental Sciences, Universidad de Cuenca, Ecuador
[2]Laboratory for Climatology and Remote Sensing, Faculty of Geography, University of Marburg, Germany
[3]Engineering Faculty, Universidad de Cuenca, Ecuador

May 5th, 2020

Why forecasting floods?

- Hydrological extremes (especially floods) have multiple impacts on society.
- Flood frequency and severity will increase with climate and land use changes!
- Forecasting is an emerging field of research for risk assessment and mitigation
- 263 floods caused more than 400 human losses in Ecuador (1970-2007)

Effect of the El Niño S.O., Ecuador, 2016.
Photo: John Sackton

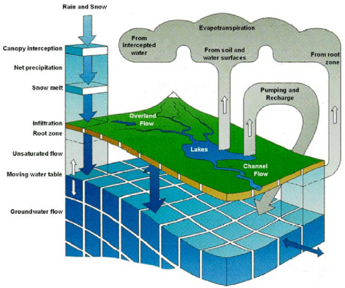
Flood forecasting in mountainous areas is harder...

- The main flood drivers (humid areas) are precipitation, soil humidity and topography.
- Flood forecasting is crucial but limited.
    - Extreme spatio-temporal variability of driving forces
    - Budget constrains
    - Remoteness of monitoring sites
- Temporal information is still limited.

Urban flash-flood in Cuenca, Ecuador, 2012.

Forecast modeling approaches

- Distributed vs. black-box modeling
- Data scarcity and model specificities (overparameterization) often limit the model operational value
- Machine Learning (ML) techniques use have increased in past decades
- ML models can deal with:
  - Missing information
  - Measurement errors
  - Non-stationary problems



MIKE SHE scheme (fully-distributed). DHI©

Research objective

Primary objective

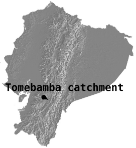
To compare the performance of five Machine Learning (ML) classification techniques for Flood Early Warning Systems (FEWSs) applied in a medium-size mountain catchment representative of the tropical Andes in Ecuador

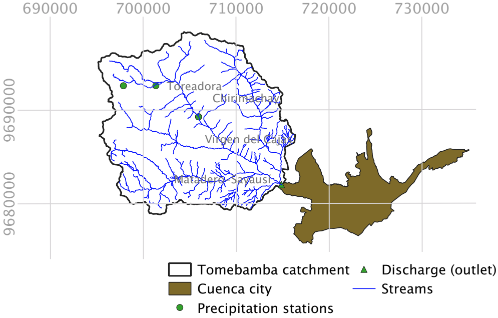
Scope:

1. FEWS alternatives are determined by the ML algorithm used and for varying lead times of 1, 4, 8 and 12 hours
2. Determination of an optimal input-structure construction process for improving forecasts
3. Determination of an appropriate methodology for performance evaluation

The Tomebamba catchment, southern Tropical Andes of Ecuador

- Area≈ 85 km^2, 2800–4400 m.a.s.l.
- Fresh water supplier of the 3rd. largest city in Ecuador (Cuenca).
- Mainly composed by páramo ecosystem.





Hydrological data

- 4 years of hourly precipitation and runoff timeseries (Jan 2015 - Jan 2019)
- Precipitation stations (within the catchment)
  - Toreadora (3955 m a.s.l.)
  - Virgen del Cajas (3626 m a.s.l.)
  - Chirimachay (3298 m a.s.l.)
- Runoff station (outlet)
  - Matadero-Sayausí (2693 m a.s.l.)
- Training: 2015-2017
- Test: 2018

ML techniques for flood classification

ML picked up from different families (classification by similarity in terms of their functionality)

1. Logistic Regression (**LR**) $\rightarrow$ Regression
2. K-Nearest Neighbors (**KNN**) $\rightarrow$ Instance-based
3. Random Forest (**RF**) $\rightarrow$ Decision tree
4. Naive Bayes (**NB**) $\rightarrow$ Bayesian theorem
5. Multi-layer Perceptron (**MLP**) $\rightarrow$ Artificial Neural Networks

Process for input construction

- Input data composition (Lag analyses)
  - Pearson´s cross-correlations for precipitation (3 stations)
  - Auto- and partial-auto-correlation functions for discharge (1 station at the outlet of the catchment)
- Feature scaling and normalization
- Principal Component Analysis (PCA)
  - Dimension reduction (trimming off correlated features)
- Resampling (under or over) ✗
- Put weights on errors proportional to class imbalance ✓

Output data labeling

- Flood warnings based on measured runoff at the Matadero-Sayausí control station
    - Control station at the entrance of the city
    - 20 years of data
- Flood warning definitions
    - Alarm $\Leftrightarrow$ $Runoff > 50\ \mathrm{m}^3.s^{-1}$
    - Pre-alarm $\Leftrightarrow$ $30 \leq Runoff \leq 50\ \mathrm{m}^3.s^{-1}$
    - No-alarm $\Leftrightarrow$ $Runoff < 30\ \mathrm{m}^3.s^{-1}$

Models' hyper-parameterization

- Each model is a combination of the ML selected and lead time (1, 4, 8 and 12 hours)
- Grid-search (10-fold cross-validation)

Model-hyper-parameters and their search domain for tuning

ML technique	Hyper-parameters				
LR	*C* 0.001 – 1000	*penalty* {'l1','l2'}			
KNN	*n_neighbors* 3 – 75	*weights* {'uniform', 'distance'}	*metric* {'euclidean', 'manhattan', 'minkowski'}	*algorithm* {'auto','ball_tree', 'kd_tree','brute'}	
RF	*n_estimators* 50 –1000	*max_features* {'auto', 'sqrt', 'log2'}	*max_depth* 50 –1000	*min_samples_leaf* 1–500	*min_samples_split* 1–500
MLP	*solver* {'lbfgs'}	*max_iter* 10 – 5000	*alpha* 1 E–9 – 0.1	*hidden_layers* 1 – 16	

Models' performance evaluation

Metrics for dealing with imbalanced and multi-class problems

- F1 score

$$F1_{score} = 2 * Precision * Recall/(Precision + Recall)$$
$$Precision = TP/(TP + FP)$$
$$Recall = TP/(TP + FN)$$

- Geometric mean

$$G_{mean} = \sqrt{TP_{rate} * TN_{rate}}$$
$$TP_{rate} = Recall$$
$$TN_{rate} = TN/(TN + FP)$$

where TP stands for True Positives, TN stands for True Negatives, FP for False Positives and FN for False Negatives

Models' performance evaluation

- Log loss score

$$Logloss_{score} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log p_{ij}$$

  where N is the number of samples, M is the number of classes, y_{ij} is 1 when the observation belongs to class j; else 0, and p_{ij} is the predicted probability that the observation belongs to class j

- Statistical significant test
  - Chi-squared test
  - To prove that the difference in the observed proportions of the contingency tables of a pair of ML algorithms are significant

Imbalanced sample class distribution

The imbalanced data problem was overcame by penalizing misclassifications inversely proportional to class frequencies:

$$w_{No-alert} = 0.01; w_{Pre-alert} = 0.55; w_{Alert} = 0.51$$

Sample class distribution for the entire dataset, and for the training and test subsets

Warning	Complete	Training	Test
No-alert	96.1%	96.2%	95.7%
Pre-alert	2.1 %	1.8 %	3.1 %
Alert	1.8 %	2.0 %	1.2 %

Lag analyses for the timeseries. 1h lead time

- Discharge
  - Dominance of the autoregressive over the moving-average process
  - 8 lags (hours) for the 1-hour lead time case
- Precipitation
  - Max correlation at lag 4 (all stations)
  - Resulting number of lags (correlation threshold of 0.2): 15 for Toreadora, 11 for Virgen and 14 for Chirimachay
- The same analyses were done for the 4, 8 and 12-hour cases

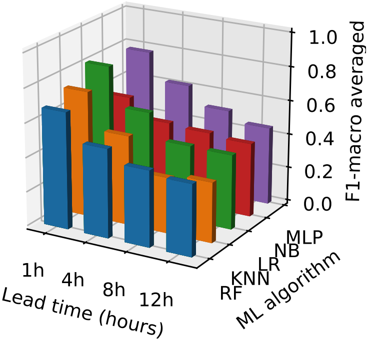
Full details on the followed methodology can be found in *Muñoz et al. (2018)*

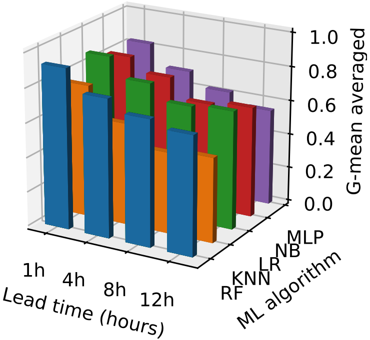
Training subset

Lead time (hours)	Performance's ranking (best → worst)				
	$F1_{macro-score}$				
1	MLP	LR	KNN	NB	RF
4	MLP	LR	KNN	NB	RF
8	MLP	KNN	LR	NB	RF
12	MLP	LR	NB	KNN	RF
	$G_{mean-macro}$				
1	RF	MLP	NB	LR	KNN
4	RF	LR	NB	MLP	KNN
8	LR	RF	MLP	NB	KNN
12	RF	LR	NB	MLP	KNN
	$Logloss_{macro-score}$				
1	MLP	KNN	RF	LR	NB
4	KNN	MLP	RF	LR	NB
8	KNN	MLP	NB	RF	LR
12	MLP	KNN	LR	RF	NB

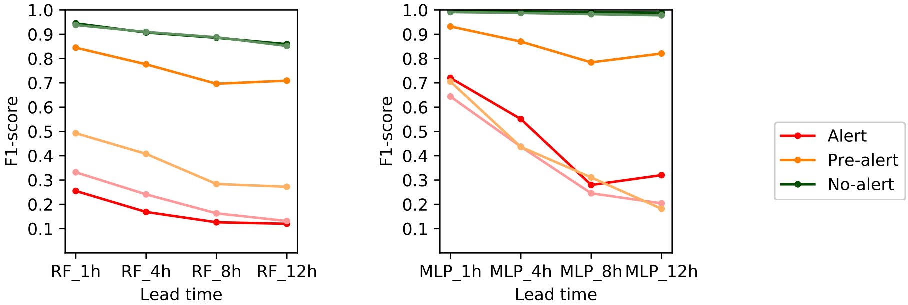
All improvements and degradations are statistically significant

Test subset


1.0
0.8
0.6
0.4
0.2
0.0
F1-macro averaged
1h
4h
8h
12h
Lead time (hours)
RF
KNN
LR
NB
MLP
ML algorithm


1.0
0.8
0.6
0.4
0.2
0.0
G-mean averaged
1h
4h
8h
12h
Lead time (hours)
RF
KNN
LR
NB
MLP
ML algorithm

A more detailed (individualized) assessment...

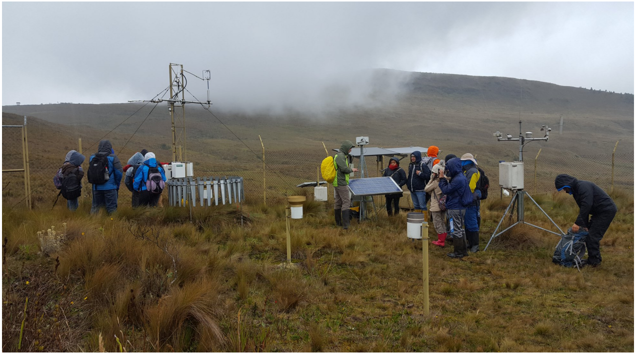


Conclusions and remarks

- The most effective model (F1-macro score and Log-loss score) was based on the MLP technique, followed by LR
- Individual F1-scores unraveled the difficulties when forecasting the Pre-alert and specially the Alert classes
- Deep exploration on the effect of input data composition and the architecture of the MLP might improve models' performance and even to extend the lead time
- The MLP can be used for the first FEWS of the city of Cuenca
- Future efforts should be put on the development of a website and/or mobile application as a tool to boost the preparedness against floods

Interested in collaborating with us?

We study the tropical Andes, the most diverse hotspot of the planet and early indicator of global change



More info (➤) Department of Water Resources and Environmental Sciences
Contact: **paul.munozp@ucuenca.edu.ec**