Random forest algorithm as a regionalization model of flood-mechanisms

Daniela Pavia Santolamazza[1,2], Henning Lebrenz[1], András Bárdossy[2]

daniela.paviasantolamazza@fhnw.ch

1. Fachhochschule Nordwestschweiz
Hochschule für Architektur, Bau und Geomatik

2. Universität Stuttgart
Institut für Wasser- und Umweltsystemmodellierung
Lehrstuhl für Hydrologie und Geohydrologie

Objective

Regionalization of extreme discharges (peak, volume and hydrograph) using **random forest** (machine learning algorithm)

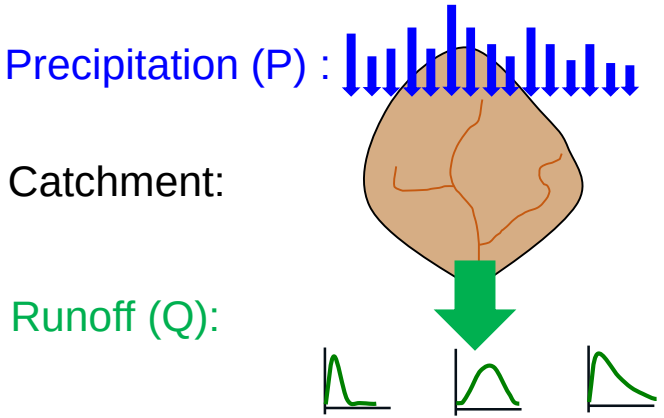
Considering:

PREDICTORS:

- **Input:** Meteorology
- **System:** Catchment hydrology and climate
- **System:** Catchment characteristics

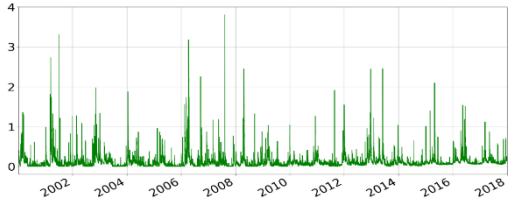
TARGET

- **Output:** Peak and volume

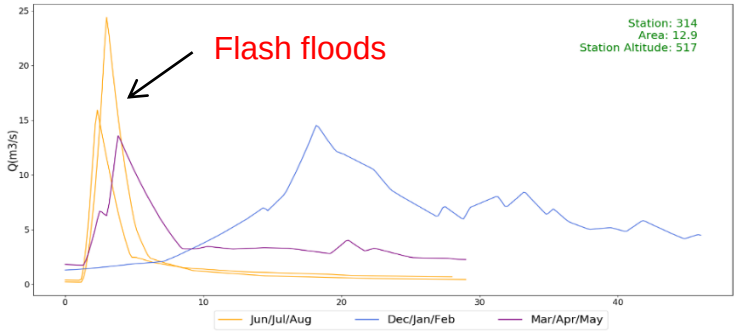


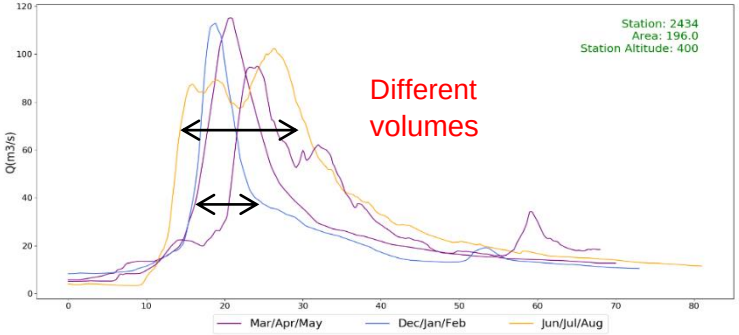
Relevance

- Length of discharge measurements short for extreme value statistics



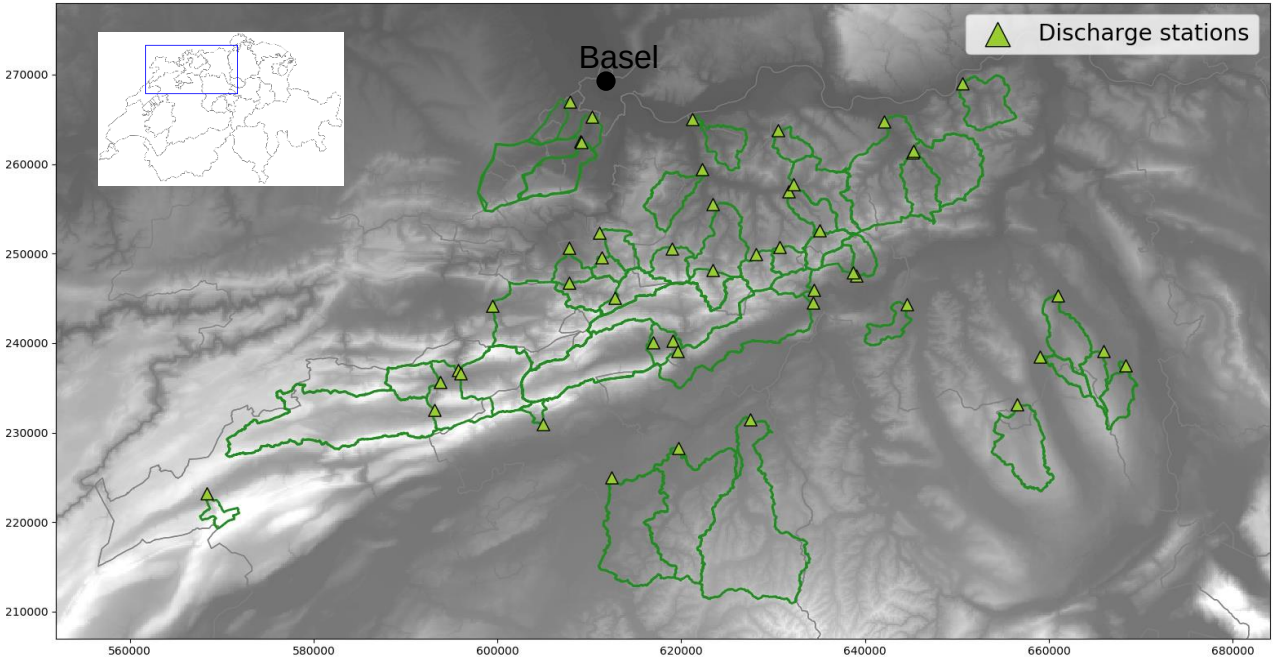
- Daily data for describing complex reactions that occur in sub-daily scales
- Commonly one random variable is used for predicting different processes





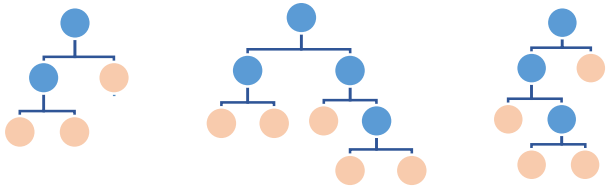
Study area

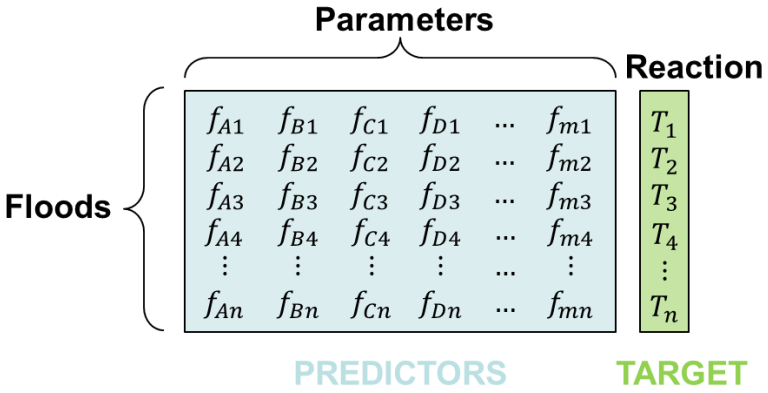
- 44 catchments with 6 km^2 to 200 km^2
- Data length: 7 - 33 years measurements every 10-15 minutes



What is Random Forest and how we use it?

*"A random forest is a classifier consisting of a collection of tree-structured classifiers {h(**x**, θ_k), k = 1, . . .} where the {θ_k} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input **x**"* Breiman (2001)





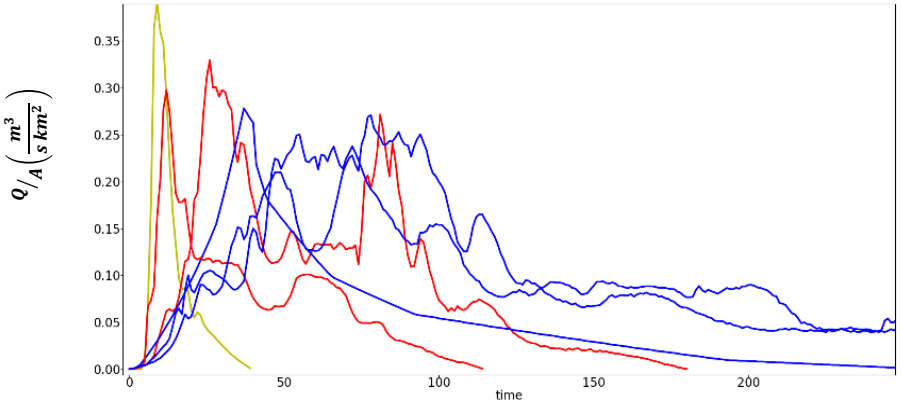
We use RF to evaluate the complex space of floods reactivity

1. Divide data series in trainset and testset
2. Train RF to predict the reaction of the catchments using the trainset
3. Use the RF model to evaluate floods proximities
4. Select possible donors
5. Estimate on testset

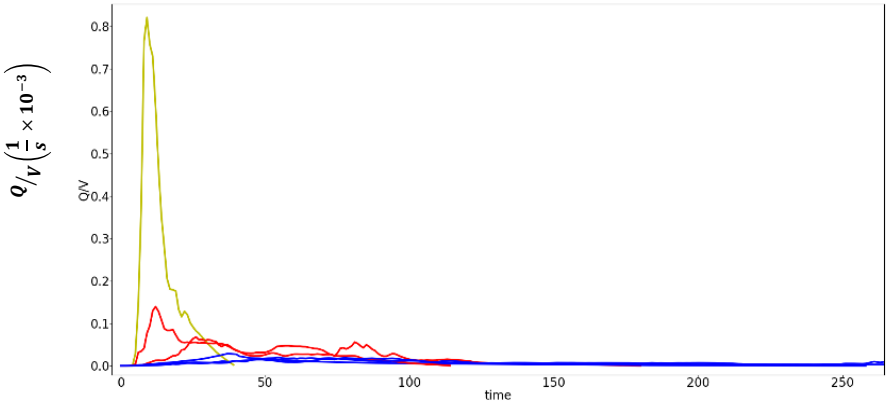
Some investigation on Random Forest

Separation of processes was achieved with a supervised RF (plots). With an unsupervised RF all floods belonging to a catchment were assigned to the same cluster.

Specific discharge



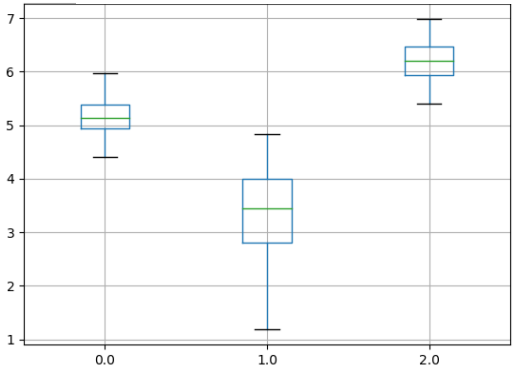
Scaled discharge



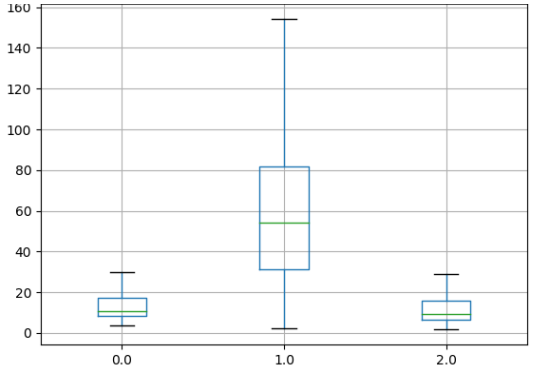
Precipitation of floods considered close by RF

Using the proximities of the RF the floods were clustered, finding that the temporal entropy as the variable with better defined groups

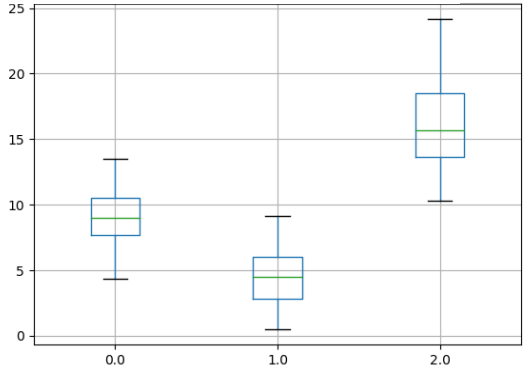
Temporal entropy



Max intensity

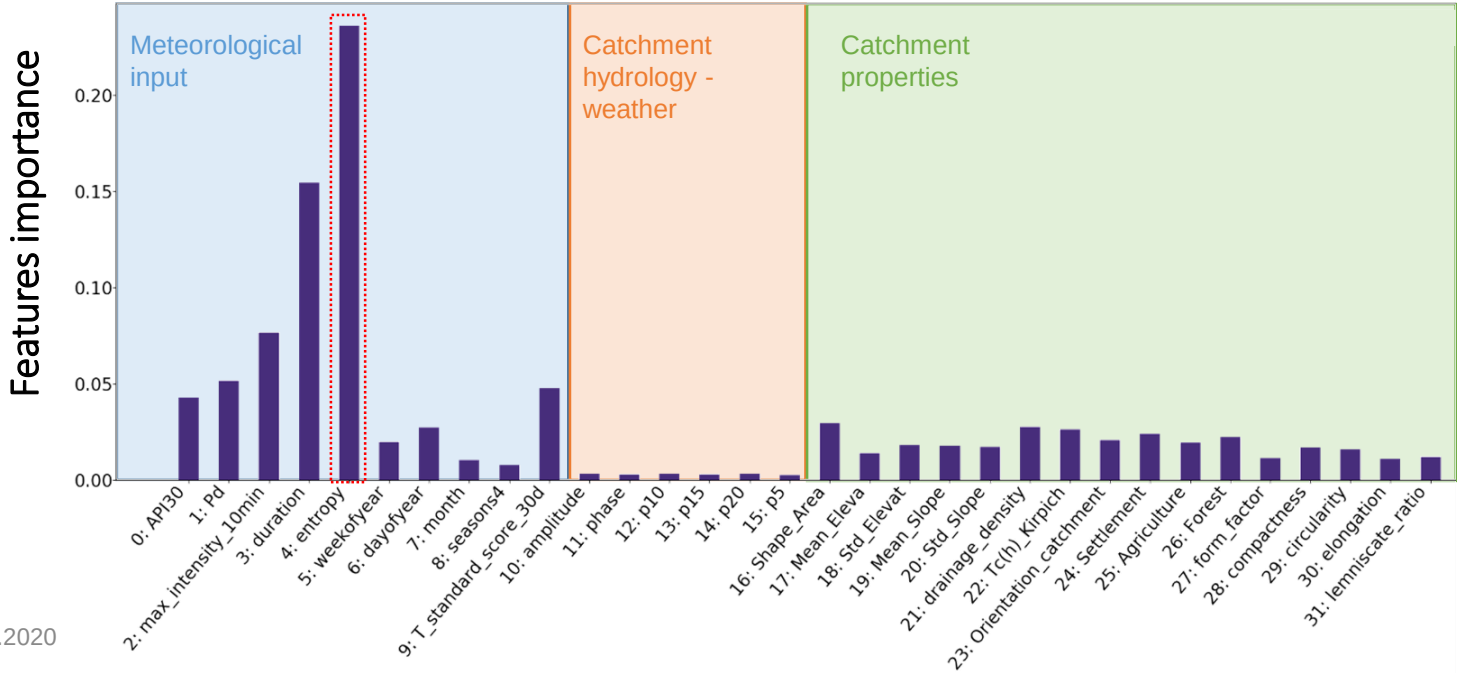


Duration



Parameters importance in RF space

Precipitation **temporal entropy** is the parameter with the highest importance on the RF model. This was the case for all models independent on target variable and covariates chosen.



What is the temporal precipitation entropy?

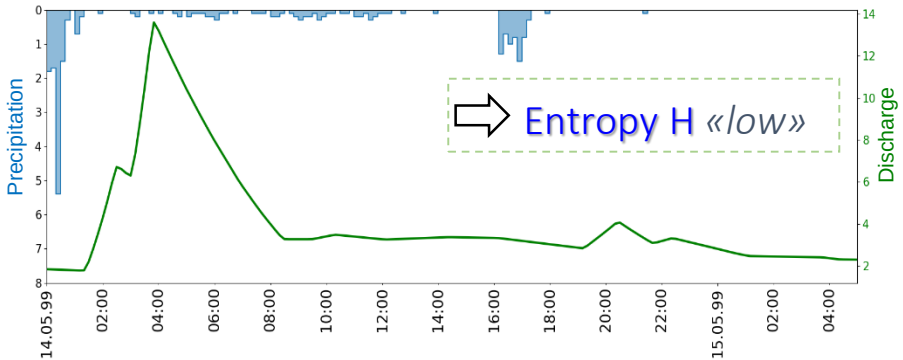
We describe temporal distribution of precipitation with Entropy H

$$\mathbf{H} = -\sum_{\mathbf{1}}^{\mathbf{i=N}} \mathbf{r_i log}(r_i) \; ; \quad 0 \leq H \leq \log(N)$$

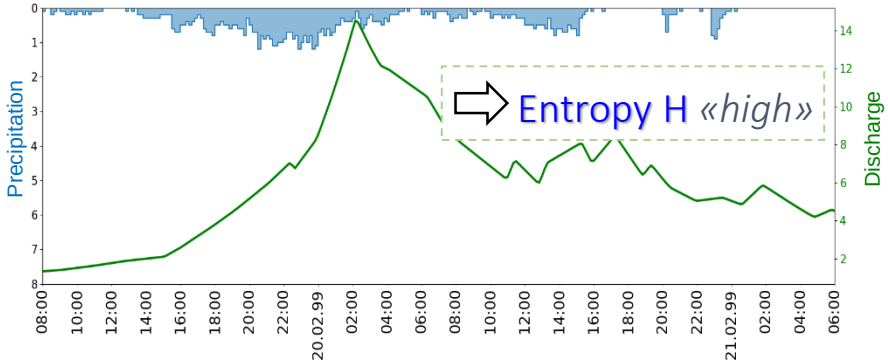
with: $r_i = \frac{P_i}{\sum_{i=1}^{N} P_i} \; ; \quad 0 \leq r_i \leq 1$

P_i: Precipitation
N: Measurements

A. **Thunderstorm** (short, intense and locally confined)



B. **Frontal event** (long, weak and widely sparse)



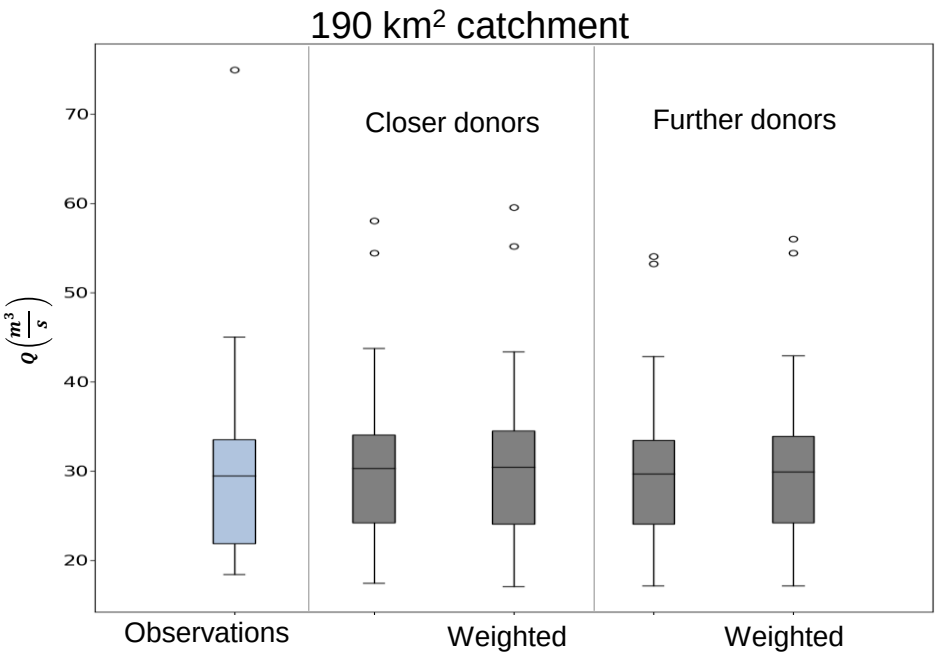
How does it behave on the ungauged catchment

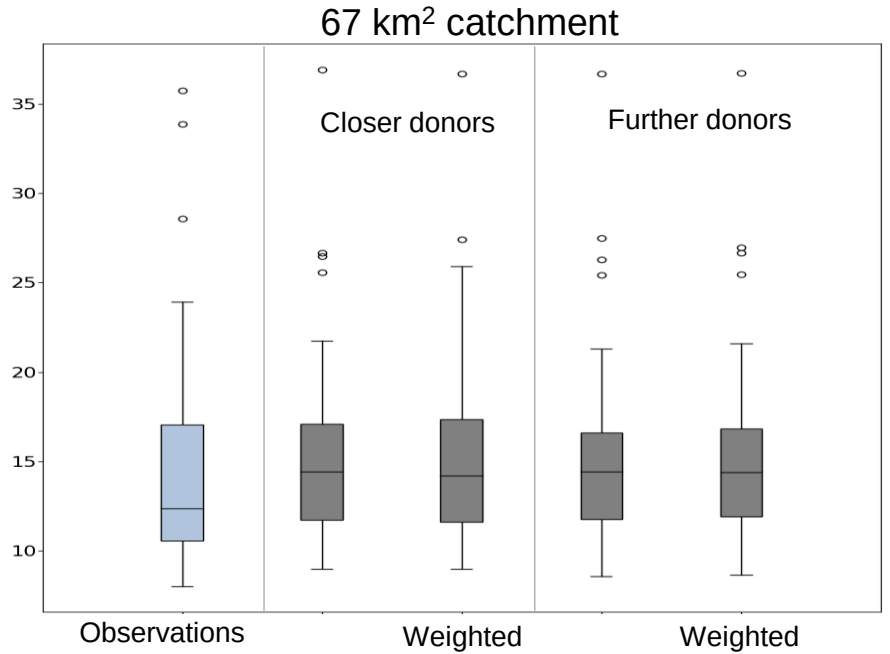
Leaving iteratively a catchment out (assumed as ungauged) peaks and peak to volume ratio were estimated.

					Nash–Sutcliffe efficiency			
					Closer donors		Further donors	
	Area (km2)	Peak/vol (1/h)	Winter floods	Summer floods	Equally weighted	Proximity weighted	Equally weighted	Proximity weighted
Mean	46	0.11	49%	51%	0.73	0.74	0.71	0.71
Median	27	0.08	50%	50%	0.78	0.78	0.76	0.76
Min	6	0.04	23%	26%	0.02	0.01	0.10	0.01
Max	208	0.40	74%	77%	0.96	0.96	0.95	0.96

Distribution of peaks

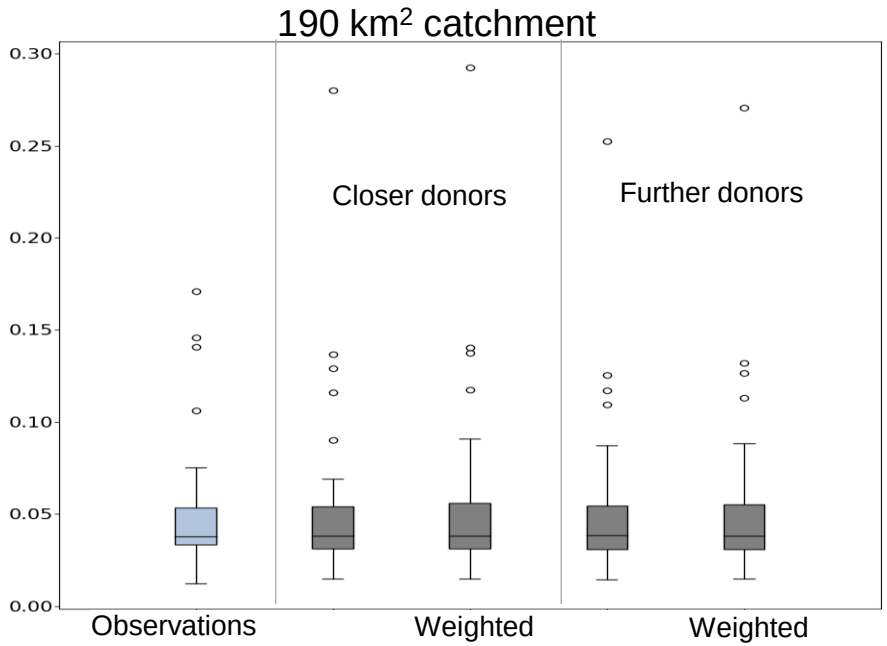
- The distribution of peaks was reproduced
- To use the RF proximity to weighting the donors improved the estimation

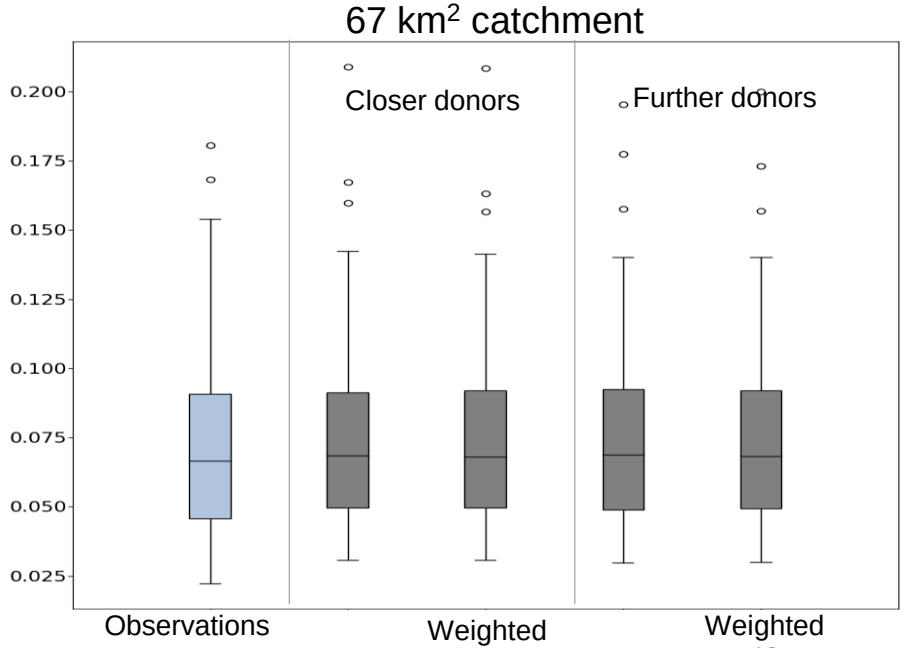




Distribution of peak to volume ration

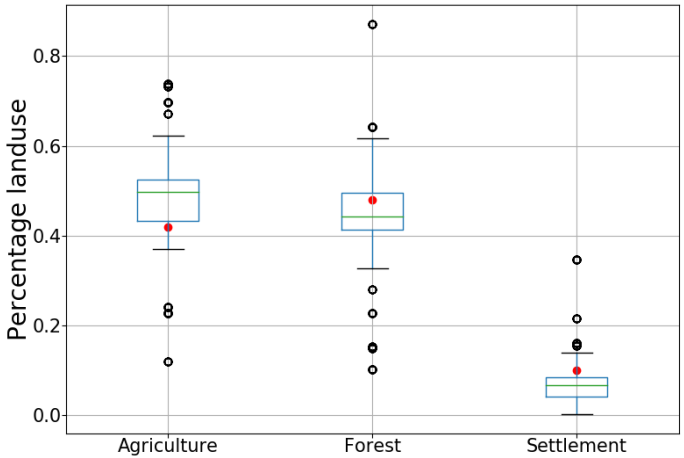
- Peak to volume ratio is a measure of the hydrograph shape
- Allowing more donors favored the estimation

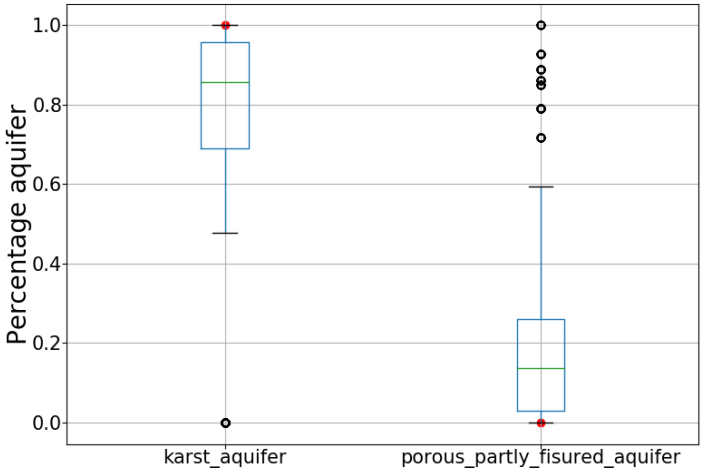




Difference of better and worse estimates

- One catchment had really low NSE. Plot placing the characteristics of the catchment in relation to the whole set, differences were found in landuse and the percentage of karst aquifer.





Conclusions

- Temporal distribution of precipitation is an important parameter for predicting hydrograph shapes
- The shape of the distributions of peaks is kept but for the higher extremes there is an over or under estimation present
- No karst parameter was consider in the RF until know. Its inclusion could help improve the estimations at some catchments
- In unbalanced data series RF estimates of the minority classes are poor. Considering some sampling techniques to equilibrate the flood processes might be helpful