Identification of flood hazard patterns over large regions using machine learning

Ricardo Tavares da Costa

17.09.2019

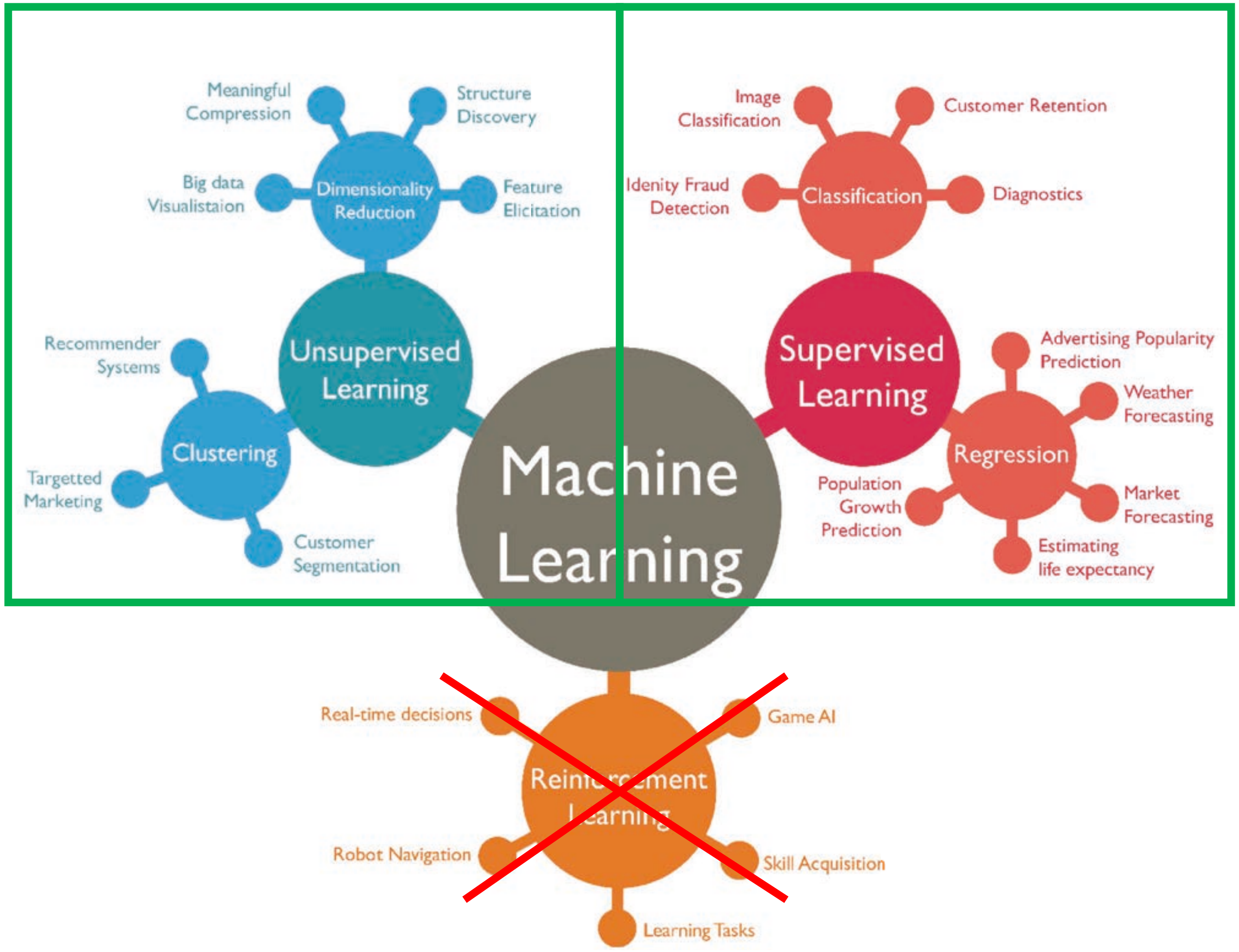
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 676027.

CONTEXT

- Floods are one of the major problems of this century
- EU Floods Directive requires EU Member States to undertake comprehensive flood risk assessments
- Flood managers, and other interested actors, lack tech options to assess flood hazard
- New techniques, such as machine learning and big open data, are not yet well exploited
- Conventional flood risk studies can be costly, time consuming, complex and can be impractical at high-resolutions or over large-scales

Machine Learning
What is?

Set of instructions (algorithms) and statistical functions that are used by computers to infer patterns from data and make predictions based on them



HYPOTHESIS

Can flood extent be regionalized?
(i.e., transferred based on a region's physical similarity/proximity)

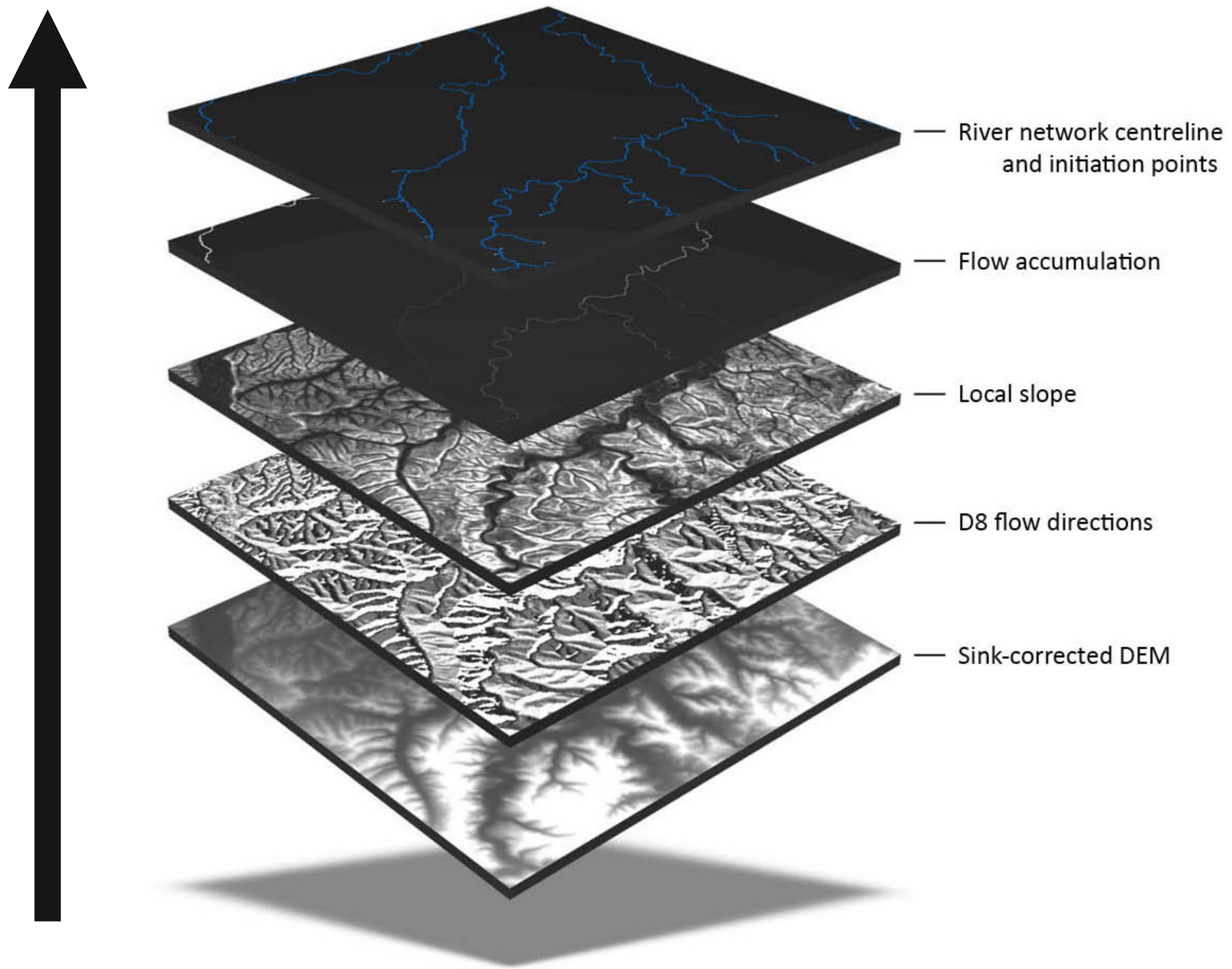
- **If** functional relations (e.g., regression) can be established between a predictor of envelope flood extent and catchment characteristics, **then** envelope flood extents can be estimated for any river basin and event likelihood.

…and why does it matter?

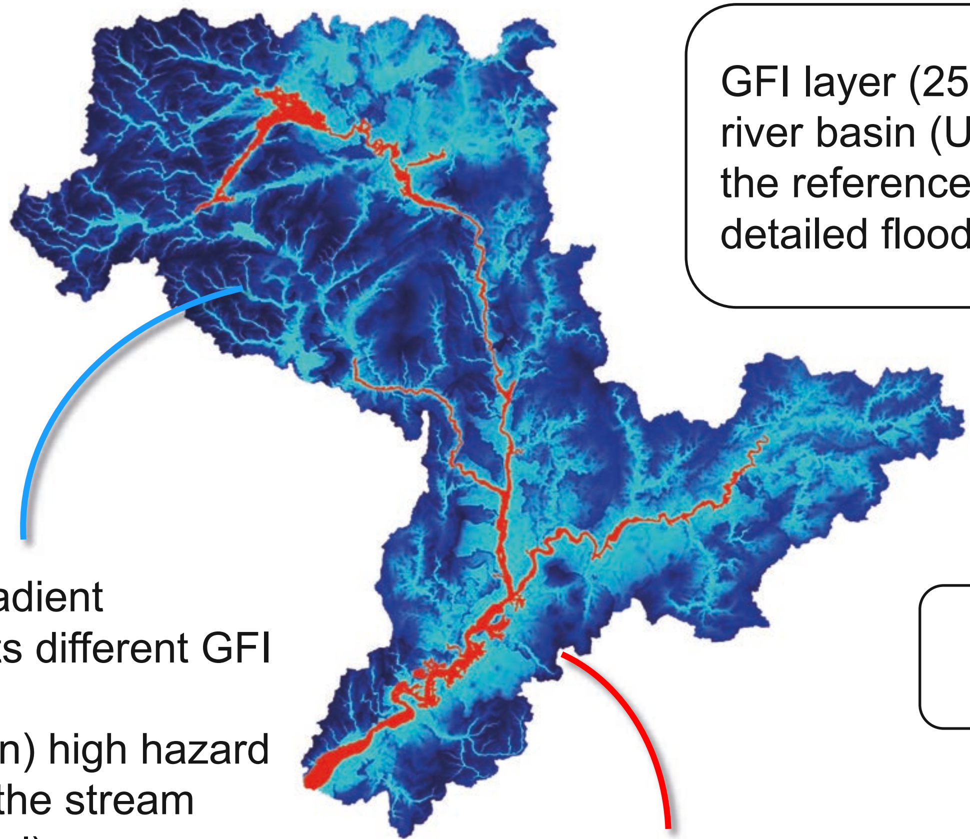
- Efficient large-scale **prediction of envelope flood extents** (including ungauged basins):

1. **Input** catchment geomorphic and climatic-hydrologic characteristics;
2. **Output** envelope flood extent for any given river basin.

PRE-PROCESSING
Terrain Analysis



CLASSIFICATION
GFI Threshold Binary Classification



GFI layer (25 m) for the Severn river basin (UK) ready to reproduce the reference flood extents from a detailed flood study

Maximize Objective Function: True Skill Score (TSS)

colour gradient represents different GFI values:

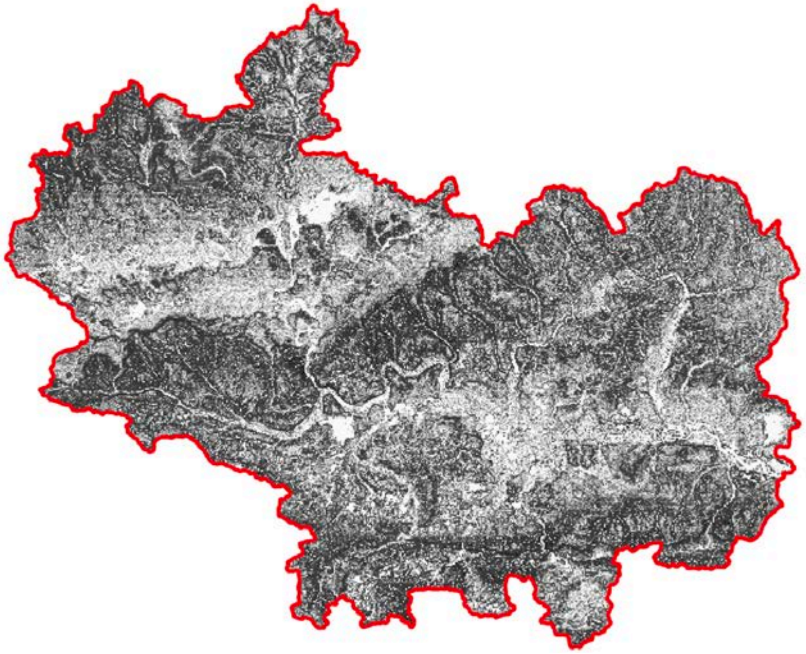
- 1 (cyan) high hazard (near the stream channel)
- 0 (dark blue) low hazard (away from the stream channel)

Benchmark flood hazard map

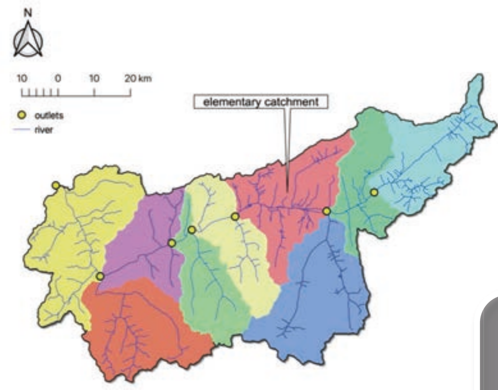
CATCHMENT CHARACTERIZATION
GEOMORPHOLOGICAL

A	Area of elementary catchment (km^2)
F	Flow accumulation at elementary catchment outlet (-)
Δz	Relief of elementary catchment (m)
S	Relief-area ratio of elementary catchment ($m\ km^{-2}$)
L_{ch}	Total river channel length in elementary catchment (km)
Δz_{ch}	Relief of the river channel in elementary catchment (m)
S_{ch}	Relief ratio of the river channel in elementary catchment ($m\ km^{-1}$)

Local slope example

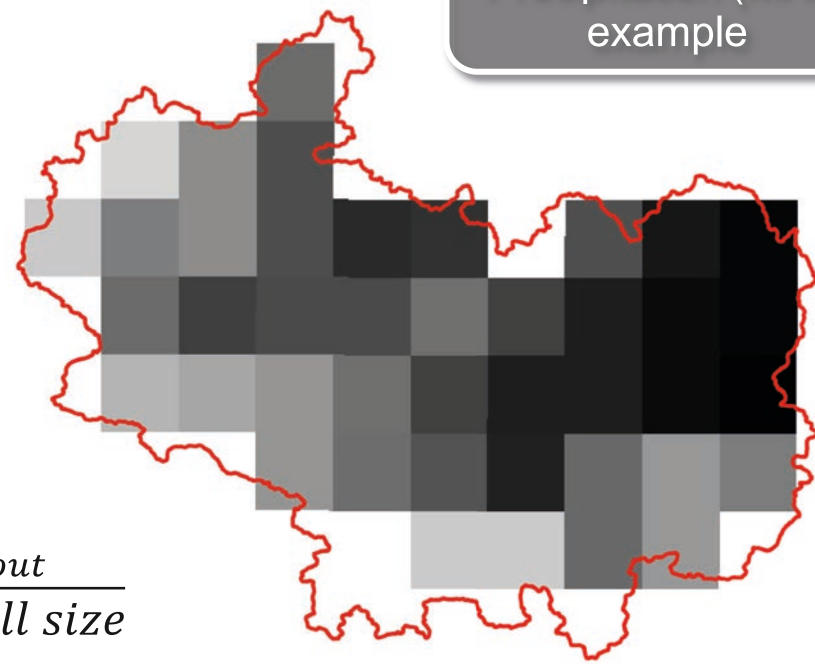


CATCHMENT CHARACTERIZATION
CLIMATIC-HYDROLOGICAL



P_{10}	10 consecutive days precipitation at elementary catchment scale associated with a 10-year return period (mm yr^{-1})
P_{10k}	10 consecutive days precipitation at elementary catchment scale, associated with a 10,000-year return period (mm yr^{-1})
MAP	Mean annual precipitation at elementary catchment (mm yr^{-1})
q_{10}	Unit discharge at elementary catchment outlet for the P_{10} precipitation statistic (m^3 s^{-1} km^{-2})
q_{10k}	Unit discharge at elementary catchment outlet for the P_{10k} precipitation statistic (m^3 s^{-1} km^{-2})
q_{MAP}	Unit discharge at elementary catchment outlet for the MAP precipitation statistic (m^3 s^{-1} km^{-2})

Mean Annual Precipitation (MAP) example



Two Return Periods!

Starting from the water balance equation:

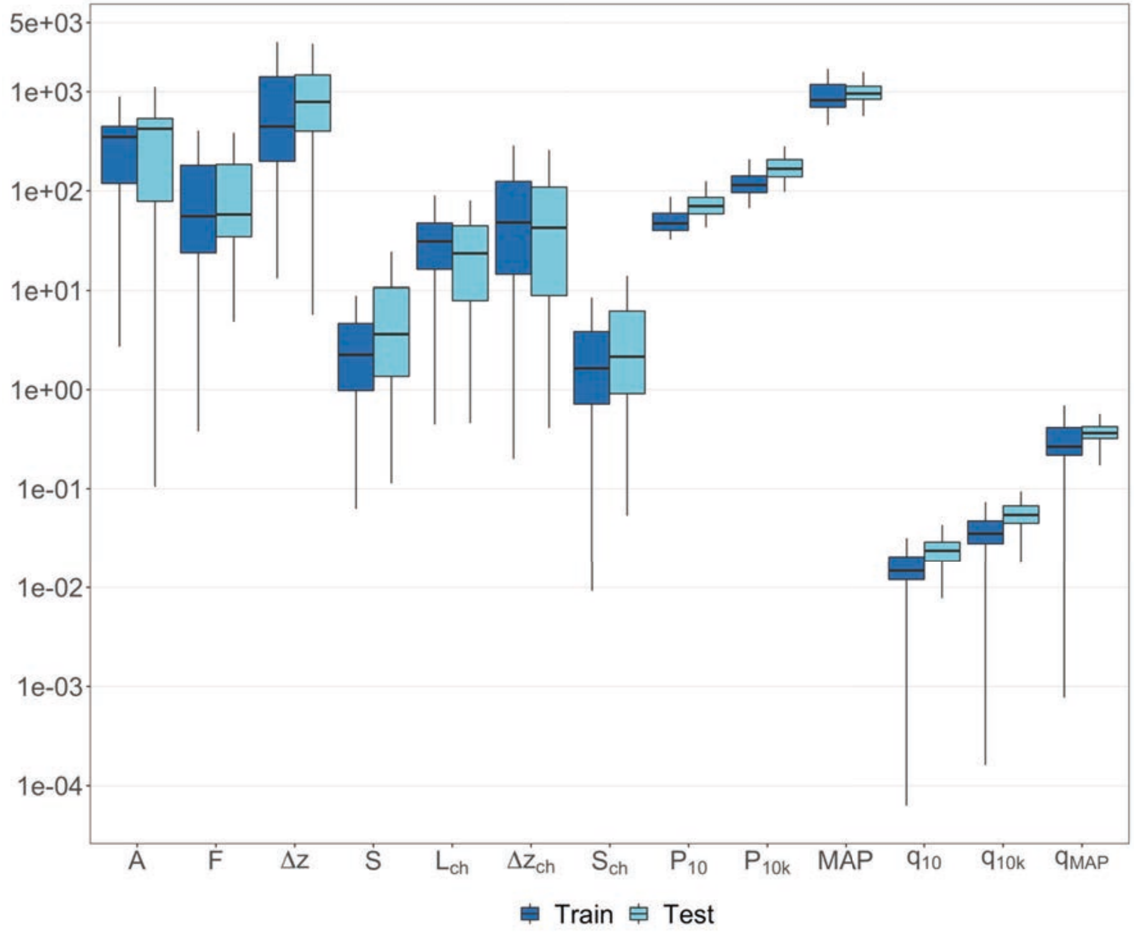
$$P + \Delta Q - E - \Delta S = 0$$

$$Q_{out} = P + Q_{in}$$

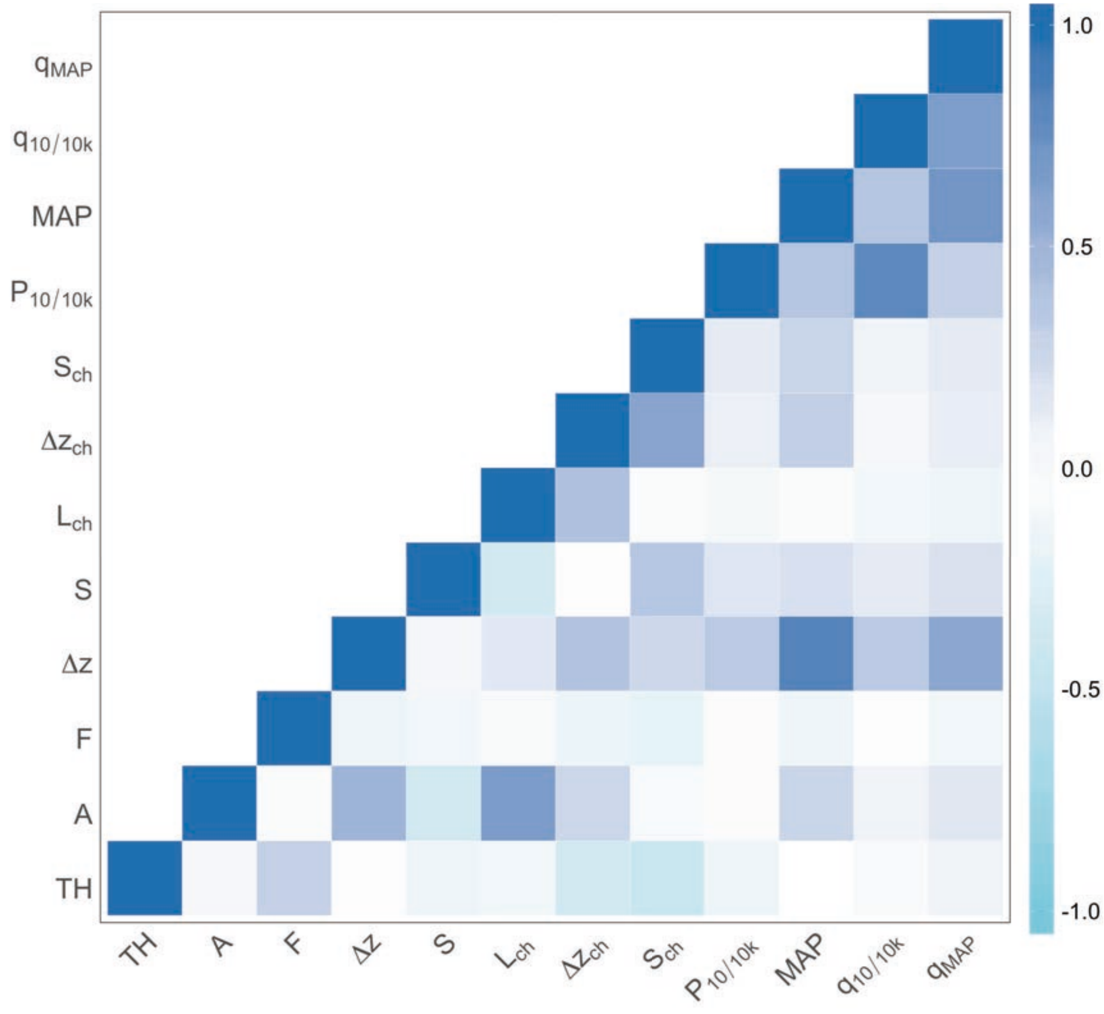
$$q_{out} = \frac{Q_{out}}{F * cell\ size}$$

$$Q_{out} = P_0 * A_0 + \sum_{i=1}^{n} P_i * A_i$$

CATCHMENT CHARACTERIZATION
OVERALL DATA DISTRIBUTION



GFI vs CATCHMENT CHARACTERISTICS
CORRELATION AND MULTICOLINEARITY



- TH and F (Flow Accumulation), moderate positive relationship.
- TH and ΔZ_{ch} (channel relief), moderate negative relationship.
- TH and S_{ch} (channel relief-ratio) strong negative negative relationship.
- TH and other characteristics, weak linear relationships.
- Correlations between catchment characteristics indicate multicollinearity.

RANDOM FOREST REGRESSION

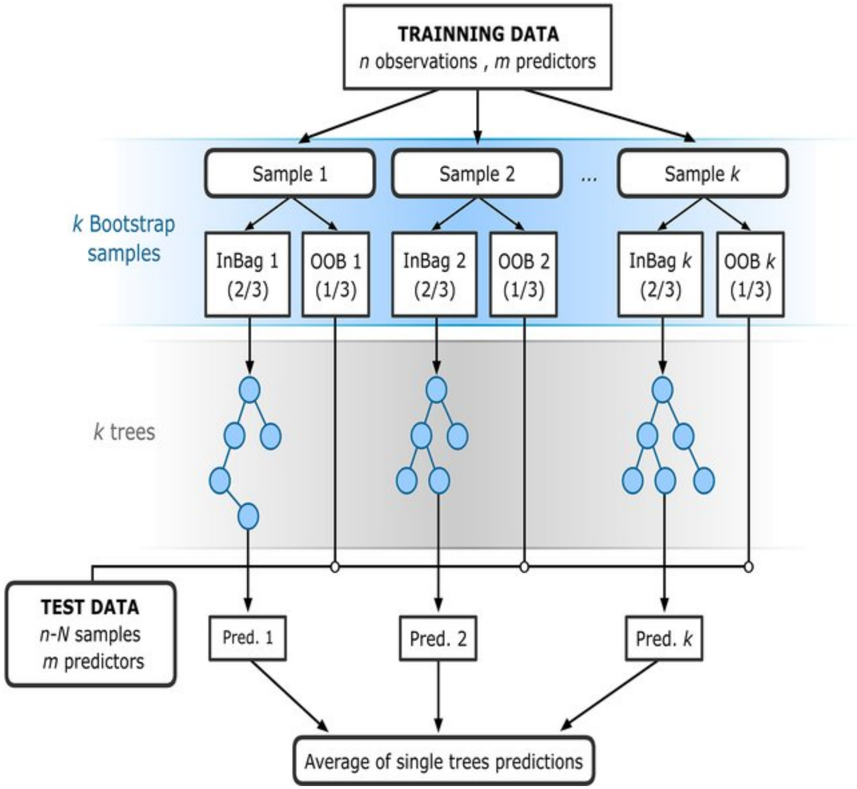
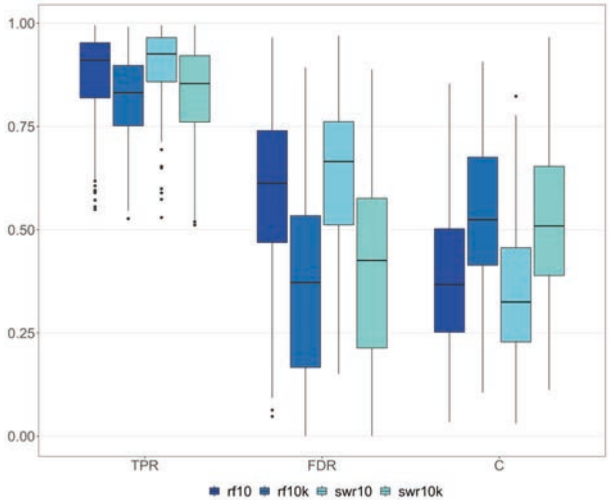


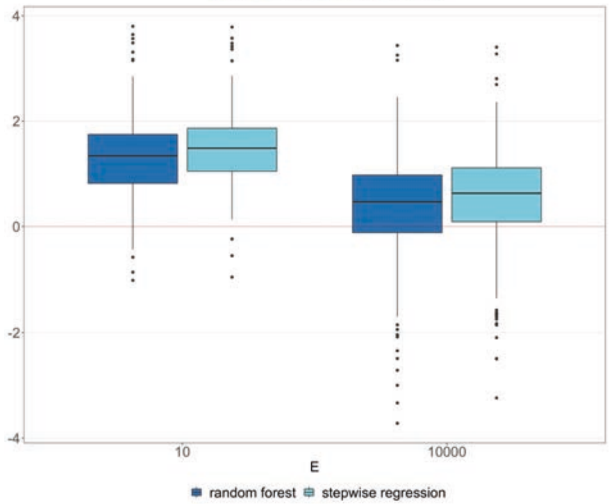
Image taken from Rodriguez-Galiano et al., 2016

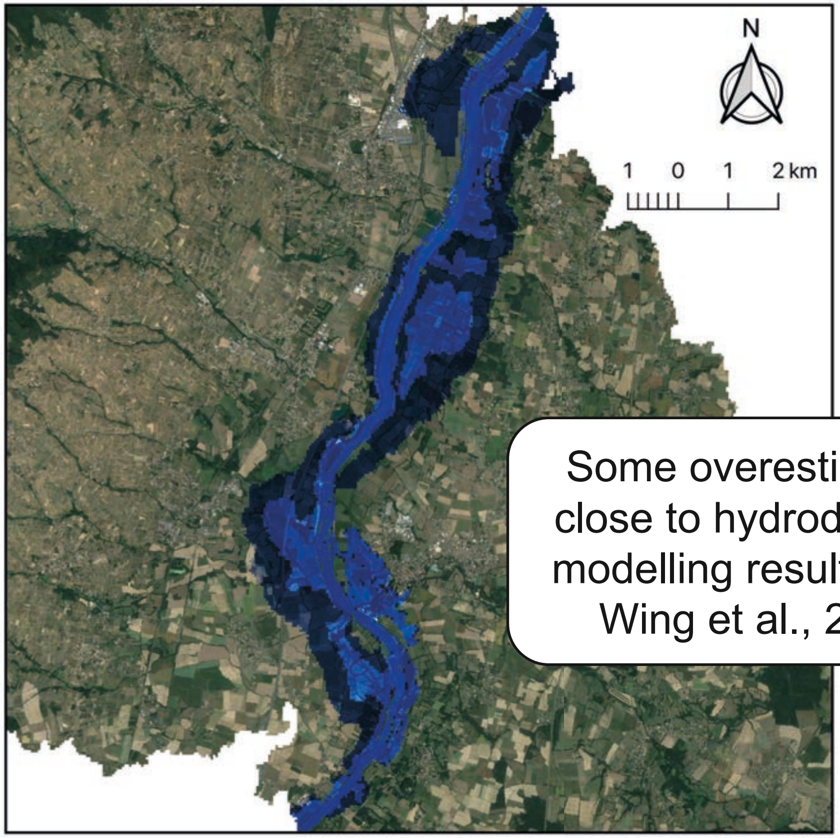
- Group of decision trees whose average output gives the final prediction.
- Random draw of samples that can be selected multiple times form independent sets that share the same distribution and form each tree.
- Nodes of a tree are data divide rules. Example which is the catchment characteristics that gives the lowest possible residual sum of squares.
- Each terminal node corresponds to a best guess of the TH.
- The tree designation comes from the hierarchy of nodes and the forest designation comes form the group of tree-like models.

RESULTS

SYSTEM RISK ETN







Some overestimation close to hydrodynamic modelling results (e.g., Wing et al., 2017)

10-year return period envelope flood extent

10000-year return period envelope flood extent

$$Hit\ Rate\ (TPR) = \frac{tp}{tp + fn}$$

$$False\ Discoveries\ (FDR) = \frac{fp}{fp + tp}$$

$$Critical\ Success\ (C) = \frac{tp}{tp + fn + fp}$$

$$Error\ Bias\ (E) = \frac{fp}{fn}$$

tp – true postives
tn – true negatives
fp – false positives
fn – false negatives

TAKE HOME MESSAGE

- This study shows that by relating classifier outcomes to catchment characteristics a less constrained mapping of flood-prone areas may be achieved for any given region, including ungauged basins.
- Prediction of flood-prone areas show that the random forest model achieves high hit rates, with average values above 60% and 80% for the 10- and the 10,000-year return periods, respectively.
- The random forest regression model more flexible and straightforward with substantially increased R^2 and decreased RMSE.
- The random forest is better suited to model non-linear behaviour and higher order interactions between catchment characteristics and the optimal GFI thresholds.
- The random forest is relatively robust against outliers, noise and overfitting and can handle the problem of multicollinearity well.

TAKE HOME MESSAGE

Limitations:

- The size and sample variability of the training set has an important impact on the performance of the approach.
- The random forest, as it is, cannot predict target values outside the range of the explanatory variables in the training dataset. This is particularly important for lower GFI values (away from the river centreline).
- The random forest does not provide an easy understanding of the statistical relationships between explanatory variables.
- The GFI underperforms specially in flat areas.

THANK YOU

Feel free to drop me a line anytime

Ricardo Tavares da Costa
ricardotavarescosta@gmail.com

FULL PARTNERS ARE:

PARTNER ORGANISATIONS:

Risk Management Solutions Ltd. (RMS)
USA

Ministry of Infrastructure and the Environment (RWS)
Netherlands

Royal Netherlands Meteorological Institute (KNMI)
Netherlands

Autorità di Bacino del Fiume Po – Po Basin Authority (AdB-Po)
Italy

Guy Carpenter
USA

Deutsche Rückversicherung AG
Germany

Landesamt für Umwelt (LfU)
State Office for Environment of the Federal State of Brandenburg
Germany

This project has received funding from the European Union‘s EU Framework Programme for Research and Innovation Horizon 2020 under Grant Agreement No. 676027

www.system-risk.eu