Predicting river flow categories using SMOS soil moisture within ML approach

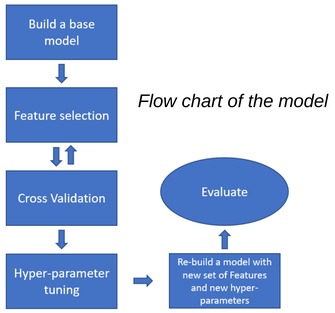
T. Jurlina[1*], C. Baugh[1], H. L. Cloke[2], C. Vitolo[1], R. Coughlan[1], F. Pappenberger[1], M. Drusch[3] C. Barnard[1], C. Prudhomme[1,4,5]

(1) European Centre for Medium-Range Weather Forecasts (ECMWF); (2) University of Reading (UoR), European Space Agency (ESA), Centre for Ecology and Hydrology (CEH), Loughborough Univerity (LU); (*) toni.jurlina@ecmwf.int

1. Introduction

We explore if the SMOS soil moisture would be a good predictor for anticipating episodes of high/low river flow. We built a machine learning ensemble classifier; Random Forest model which uses nine dynamic and four static Features for predicting river flow categories from dry to wet. The model is designed to give predictions 1, 3, 5, and 10 days in advance. Model is verified on over 608 European catchments during the period March 2017 to May 2018. Although the primary study goal was to predict river flow categories from dry to wet, the model could be used for predicting flood susceptibility.

2. Data and Methods



Flow chart of the model

Main static Features	Main dynamic Features
Catchment area	Daily SMOS soil mositure
Catchment slope	Daily precipitation
Catchment DEM	Daily temperature
Catchment channel length	Daily snow melt
Catchment concentration time	
Catchment water body fraction	
Catchment soil type	
Catchment land use type	
Catchment Koppen climate class	

Basic set of Features used for building the Multiclass classifier Random Forest model for predicting river flow categories

Discharge percentile	Joint river flow class
<20th	Very low
20th – 40th	Low
40th – 60th	Normal
60th – 80th	High
>80th	Very high

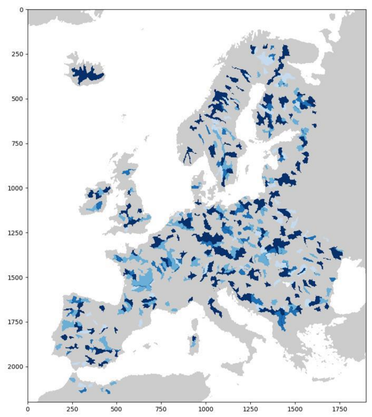
Definition of river flow categories

Machine Learning model – methodology and Features selections

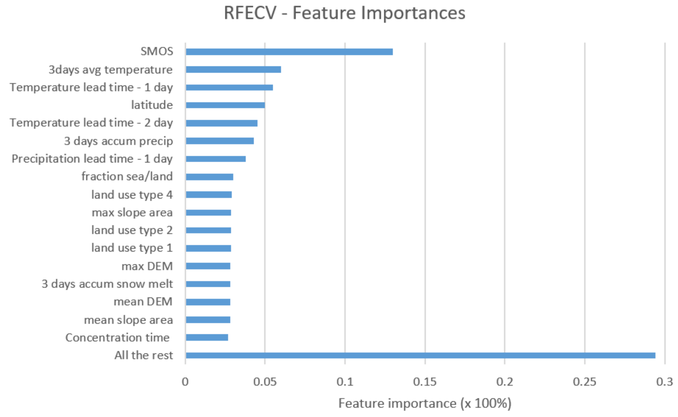
- The chosen machine learning method in this study is the Random Forest – Multiclass Classifier. It requires datasets of predictors (Features) and a predictand (labels)
- The predictand is the river flow category which we define by climatological percentiles, from low to high.
- Features that can be assumed not to vary in time (static F.), such as time of concentration or catchment elevation, describe catchment properties that are time independent. Dynamic Features, such as precipitation and temperature bring temporal variation into our system.

3. SMOS soil moisture importance

- Study is focused on pan-European domain, and the research is conducted at a catchment level. Only upstream catchments of less than 2000 km^2 are selected
- The Feature importance was calculated during the Recursive Feature Elimination (RFE) process.
- Analysing the results from the 3-day lead time model showed that the SMOS soil moisture Feature was the most important Feature in our 3-day model

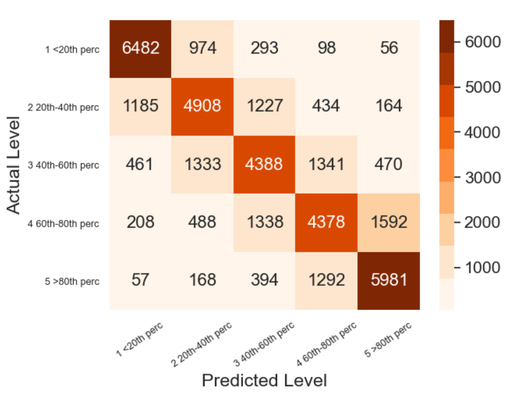


Study area with catchments smaller than 2000 km^2



Model Feature importance for the 3-day Random Forest model

4. Model evaluation

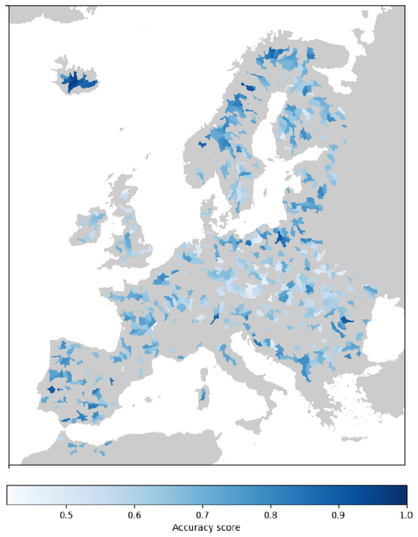


Confusion matrix, where predicted vs observed flow level classes are on each axis, for the 3-day model. Values in squares show the number of correctly predicted events

Model evaluation

- The overall accuracy scores for each lead time were greater than 0.50. The score also increased with the lead time.
- This trend could be explained because the longer lead time models have more dynamic input Features than the shorter lead time models
- It would be expected that the score should decrease with increasing lead time owing to the skill decrease of dynamic inputs such as precipitation at longer lead times. In this study however we have chosen to use observations of the dynamic Features to act as proxies for a forecast.
- The greatest values occur in the long diagonal which represents correct prediction of the different river flow categories. The values in each cell of the confusion matrix represent the total count in each cell. In the 3-day model, the greatest values occur in the two most extreme classes

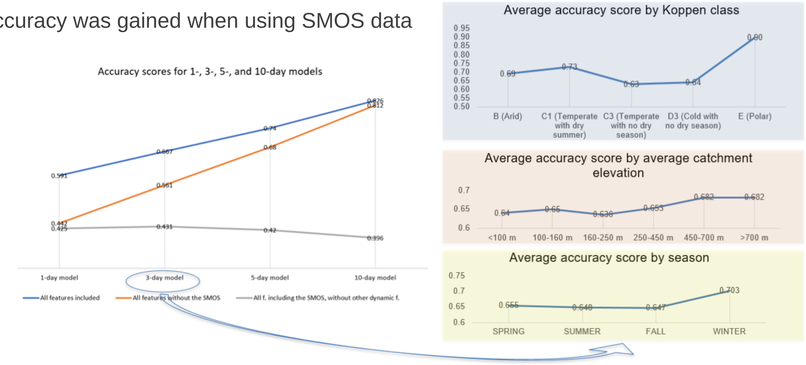
5. Spatial and seasonal model performance analysis, lead time model analysis



Accuracy score per European catchments from the 3-day lead time model

Spatial analysis

- Greatest accuracy scores are present in Sweden, Iceland, northern Poland, Spain and the Balkans. The lowest scores occurred in the UK, Ireland, central Poland, Finland and eastern Germany.
- This spatial distribution suggests that greater skill could occur in high latitude catchments which could be influenced by snowmelt as well as arid areas such as Spain
- Accuracy was gained when using SMOS data



Relative importance in accuracy score of SMOS vs no SMOS models (left), and accuracy score dependence of Koppen class, elevation, and season for the 3-day model (right)

Summary

- The soil moisture data from SMOS was the single most important Feature in each of the models that was constructed. The importance of soil moisture was greatest at shorter lead times, which is expected as the soil moisture correlation is higher from day to day, than it is for longer period.
- Although the primary study goal was to predict river flow categories from dry to wet, the model could also be used for predicting flood susceptibility when the highest flow category is predicted in a catchment. As the model performs better for the extreme river flow classes we conclude that the model is suitable for predicting flood susceptibility.
- Future work could include for a greater number of predictand flow categories which could allow for more refined predictions of flood susceptibility