HS8.2.2. – Groundwater in a changing environment. Sustainability and adaptative management of resources.

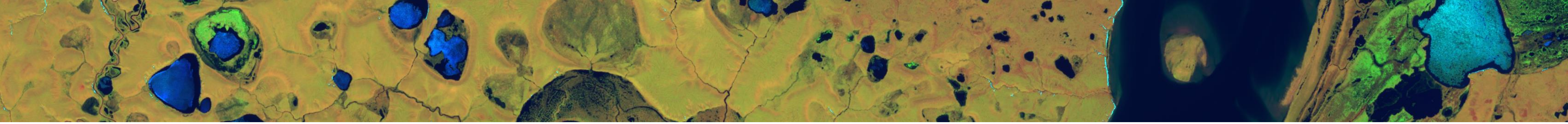
Predictive modelling of groundwater nitrate pollution at a regional scale using machine learning and feature selection

Aaron Cardenas-Martinez[1]
Victor Rodriguez-Galiano[1]
Juan Antonio Luque-Espinar[2]
Maria Paula Mendes[3]

1 – Physical Geography and Regional Geographic Analysis, University of Seville, Spain
2 – Instituto Geológico y Minero de España (IGME), Granada, Spain
3 – Civil Engineering Research and Innovation for Sustainability (CERIS), Instituto Superior Técnico, Universidade de Lisboa, Portugal

EGU General Assembly 2020

US
Universidad de Sevilla
· 1505 ·



1. Introduction

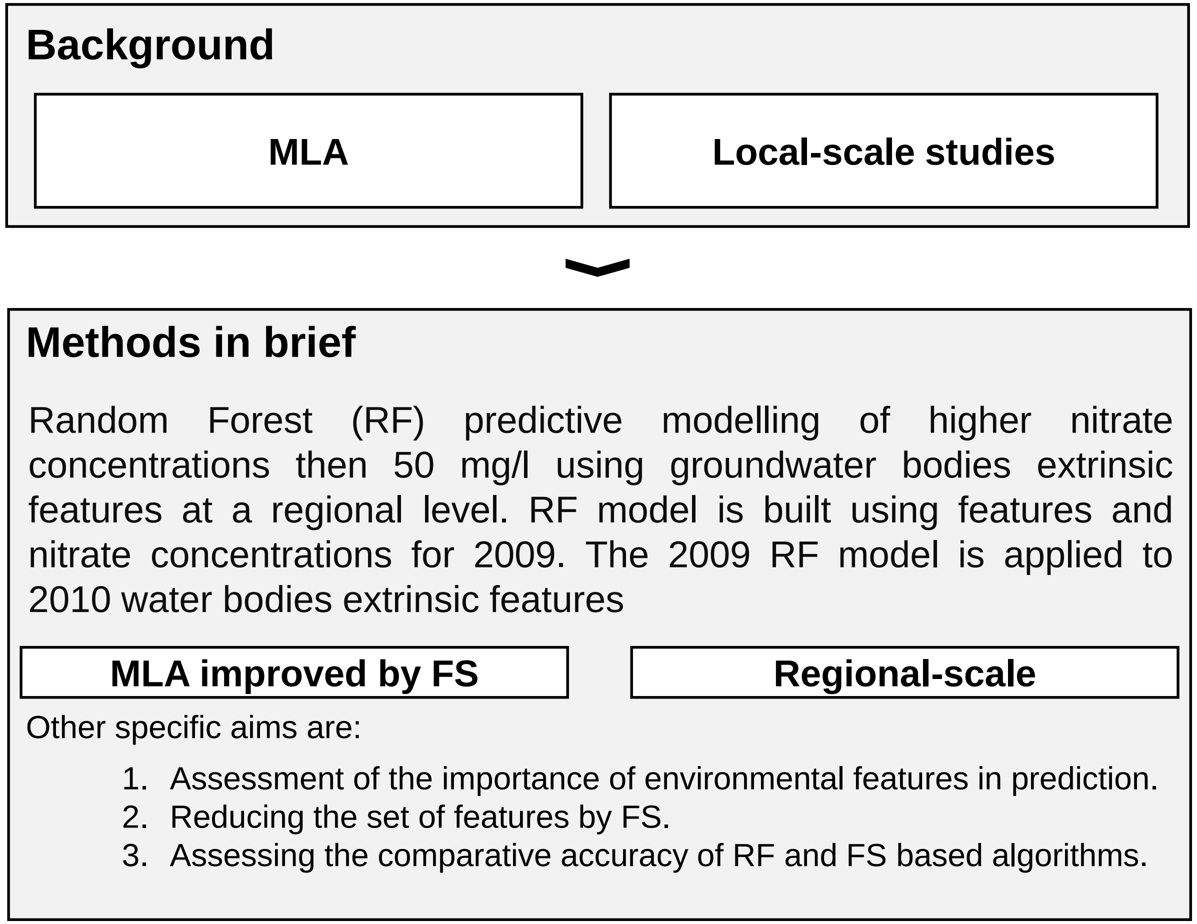
Protecting strategic resources such as water is one of today's main challenges, due to the effects that altering their quality generates on society, environment and economy. Global population growth and economic transformations have caused major changes in methods of food production, requiring nitrogen fertilizers to boost the crops.

The surplus of fertilizers especially affects groundwater bodies, which are under the influence of irrigation, shallow depth, livestock effluents and soil permeability.

The European Commission establishes through the WFD that the nitrate content of groundwater bodies should not exceed 50 mg/l. However, between 2012 and 2015 this limit was exceed by:

- 13.2% of the groundwater stations in the EU.
- 21.5% in Spain.
- 34.9% in Andalusia (Spain).

Knowledge of the concentration and distribution of nitrates in groundwater bodies is essential for their management. There are approaches to predicting the spatial distribution of nitrates:



Background

MLA

Local-scale studies

Methods in brief

Random Forest (RF) predictive modelling of higher nitrate concentrations then 50 mg/l using groundwater bodies extrinsic features at a regional level. RF model is built using features and nitrate concentrations for 2009. The 2009 RF model is applied to 2010 water bodies extrinsic features

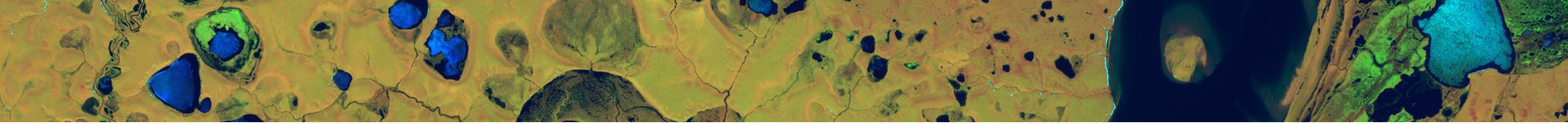
MLA improved by FS

Regional-scale

Other specific aims are:

1. Assessment of the importance of environmental features in prediction.
2. Reducing the set of features by FS.
3. Assessing the comparative accuracy of RF and FS based algorithms.

cc BY



2. Study area

The region of Andalusia is located in southern Spain and is the most populated administrative region of the country, with almost 9 million residents. Although it has a rugged terrain, it has large areas of irrigated croplands in the river basins. In 2013 there were 1,036,060 ha of irrigated crops, most of them in the Guadalquivir basin, especially in the closest area to the coast, where vast rice fields are located. Other areas with an important presence of irrigated crops are the Guadalete basin or the Vega de Granada, especially for herbaceous and olive plantations. Along with this, Campo de Dalías is the main area of greenhouse crops.

168 groundwater bodies in Andalusia were identified in the Hydrological Plans for the period 2009 - 2015 were used for this research.

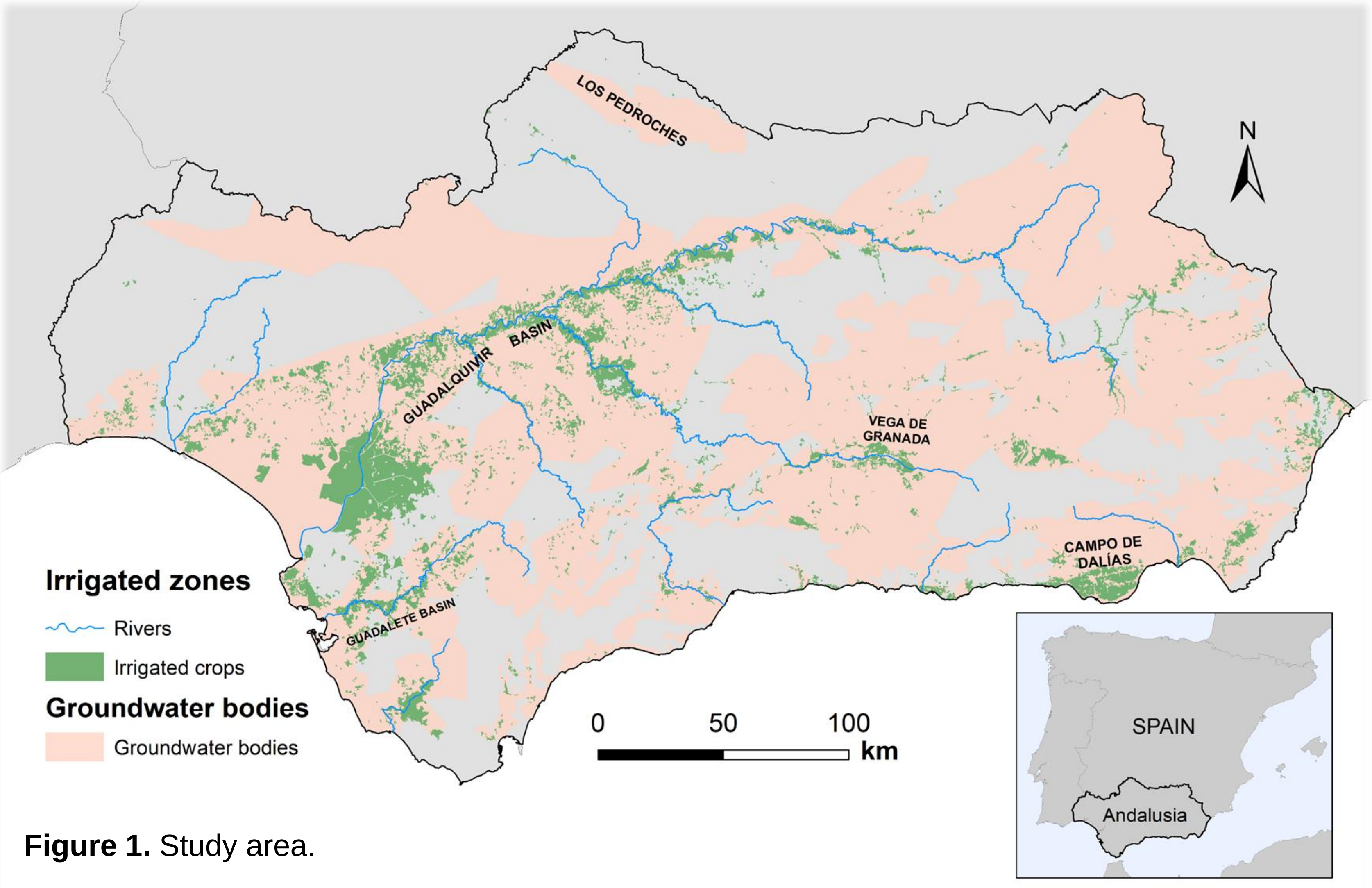
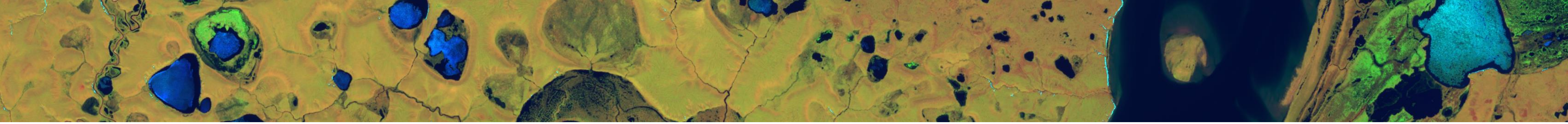


Figure 1. Study area.

cc BY



3. Methodology

Data

Nitrates data for training and prediction were obtained from a database of 4,659 samples measured between 2002 and 2010. Initially, data were filtered by month to obtain the two months of the year with the most records: June and October, to measure the changes produced by fertilization in the summer.

Fig. 2 shows the number of samples by month and year of campaign and nitrate concentration values in mg/l for each campaign. This analysis allowed us to observe that the change in the number of samples causes great variability in the dispersion of the values. 2009 and 2010 were selected since there were enough samples for the stats to be representative, joining the measures of June and October of each year to obtain two databases: one for training with 321 samples and the other with 281 samples for prediction.

Outliers, defined as 1.5 times the interquartile range, were shown to be associated with irrigated crops, so it was decided to keep them.

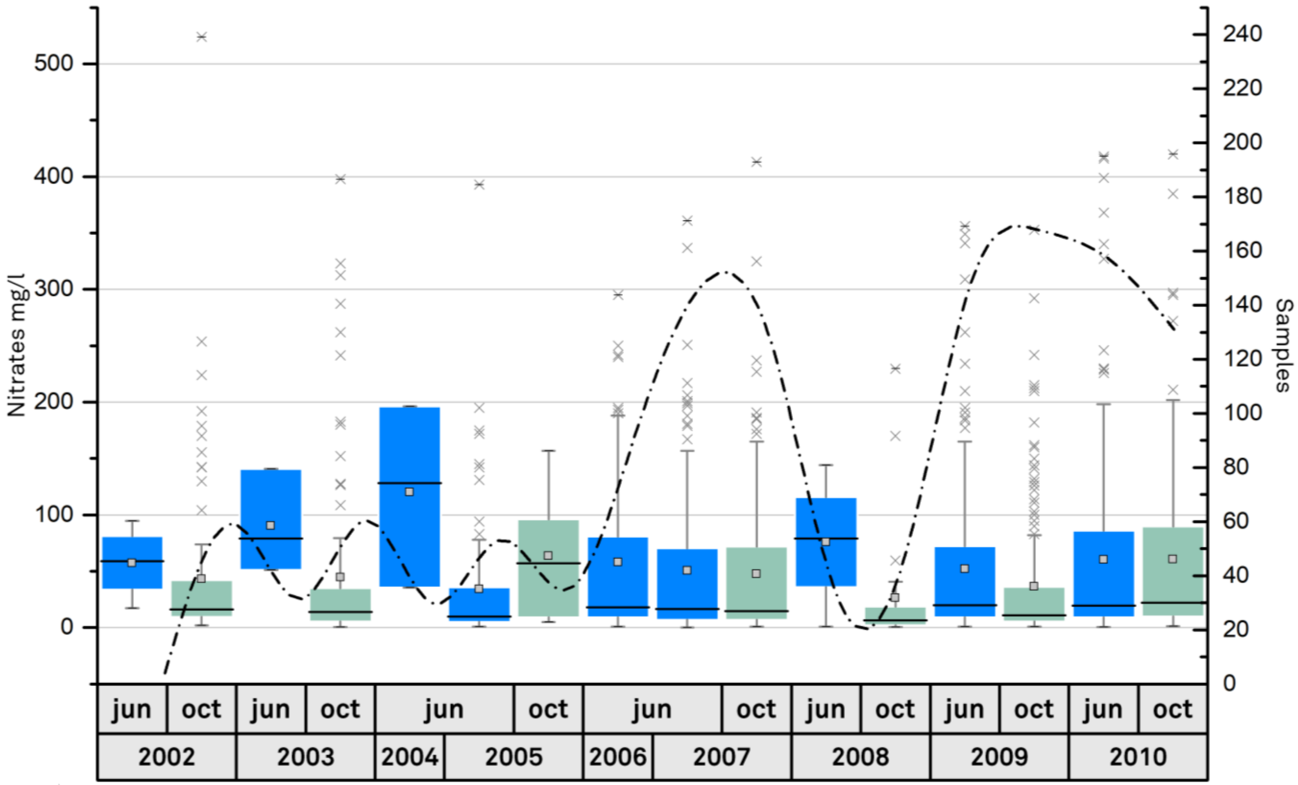
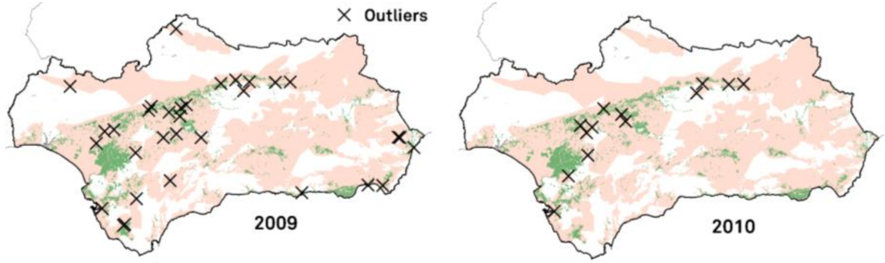
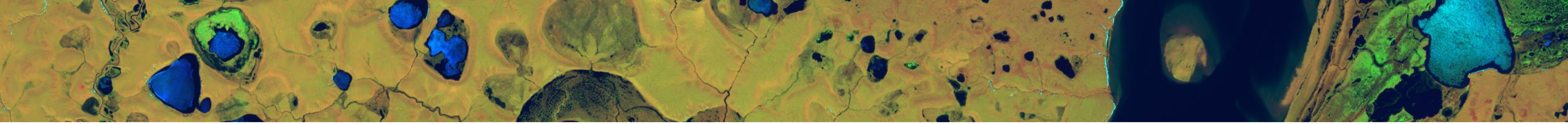


Figure 2. Box-plot of the Nitrate Database for the months of June and October. Years 2002 - 2010. The figure shows two data sets on the ordinate axis. On the one hand, the box-plot represents the nitrate concentration values in mg/l. On the other hand, the flashing line refers to the number of samples available in each campaign.

Outliers:



cc BY



3. Methodology

Database design and modeling

A set of 44 features were used to train the models. The main nitrate producers were considered: agriculture and livestock, for which 13 phenological features based on weekly MODIS reflectance compounds were used. The average and accumulated index of livestock effluents were obtained for 3 radius of action. In addition, features related to nitrate distribution and accumulation were introduced, such as 18 features related to terrain attributes, 4 to climatology and 3 to texture.

300 classification models were trained in RF by performing a hyper parameter tuning, with a sequence of 1:30 features and 100-1000 trees every 100. The models were trained with 2009 data, using 2010 for prediction and validation. As they have many predictive characteristics, an FS was applied, which helps to reduce the dimensionality, accelerate the learning process and improve the interpretation capacity. Four different methods were applied: Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), Sequential Floating Forward Selection (SFFS) and Sequential Floating Backward Selection (SFBS).

For the model validation, a 10-fold Cross-Validation was used, which allows the algorithm to learn from the entire data, being a separate procedure. For the comparison between RF and FS, the Mean Misclassification Error (MMCE) was extracted.

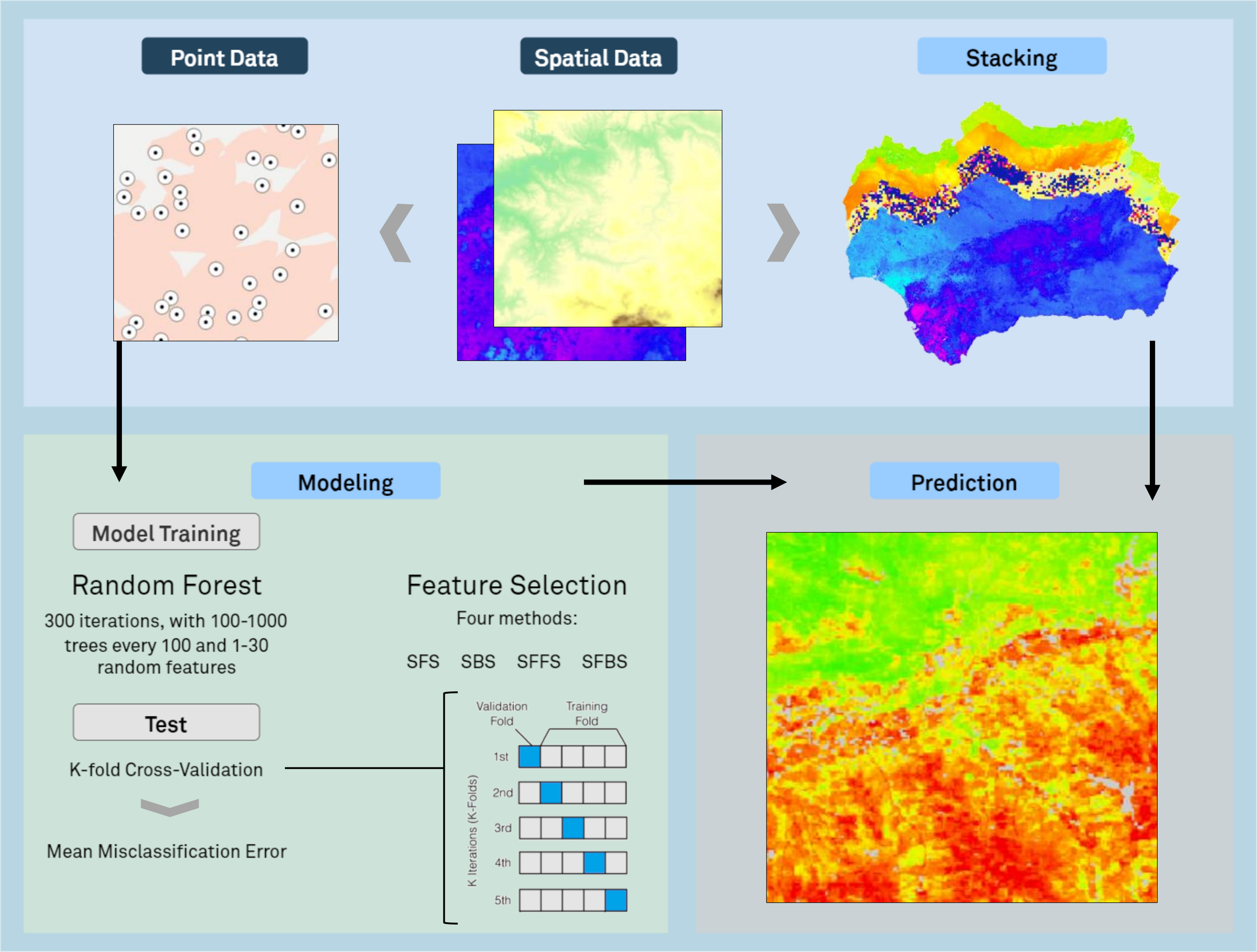
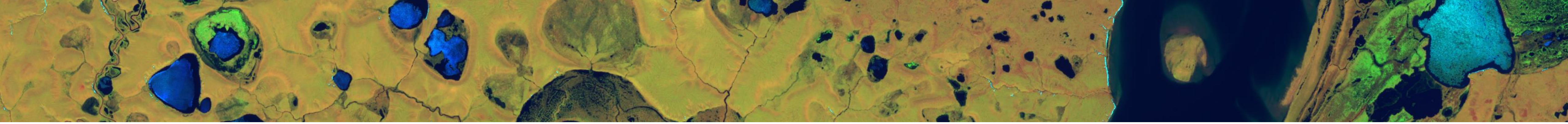


Figure 3. Work-flow. Figure shows three main steps (1) data processing; (2) modeling; (3) prediction

cc BY



4. Results & discusión

Feature importance and prediction

The results of the best model obtained in RF and FS were analyzed. The RF model applied all the features for prediction (the 15 most important are shown), while the FS model predicted with 6 features. Terrain attributes were the features that most influenced the prediction, along with phenology features. The Multi-resolution Ridge Top Flatness (MRRTF) and Multiresolution Index of Valley Bottom Flatness (MRVBF) were among the most important. Of the phenology features, the value of NDVI for the start of the season (Start Season val. NDVI), understood as the value of the function at the start date of the season. This feature indicates the value of NDVI at the beginning of the growing season which, in the case of crops, is usually very low, since they are in the germination phase. The analysis revealed that where the NDVI at the start of the season was lower, nitrate concentrations were higher.

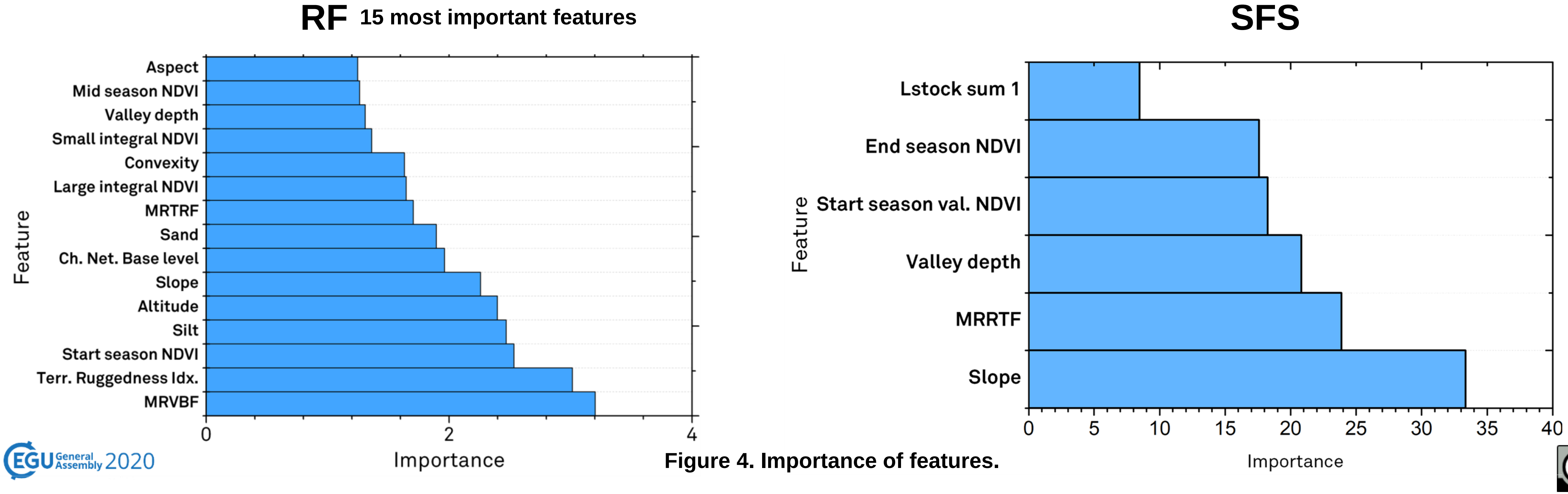
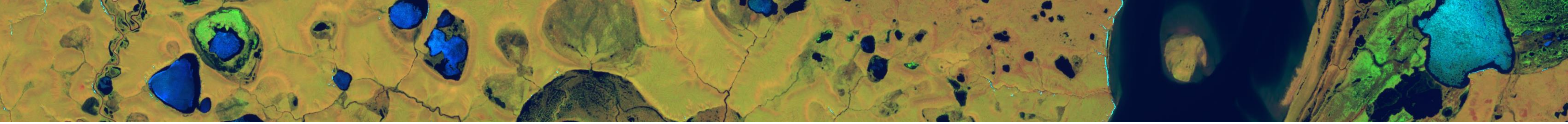


Figure 4. Importance of features.



4. Results & discusión

Feature importance and prediction

Prediction results showed in both models that the highest probabilities of finding high concentrations of nitrates were found in the Guadalquivir Basin, especially in the area where the river valley is wider. High probabilities were also found in the Guadalete Basin. The higher concentrations of nitrates in the lower sections of the rivers could be explained by the process of leaching nitrogen from fertilizers, by the action of precipitation and the return of irrigation, which contribute high amounts of nitrites.

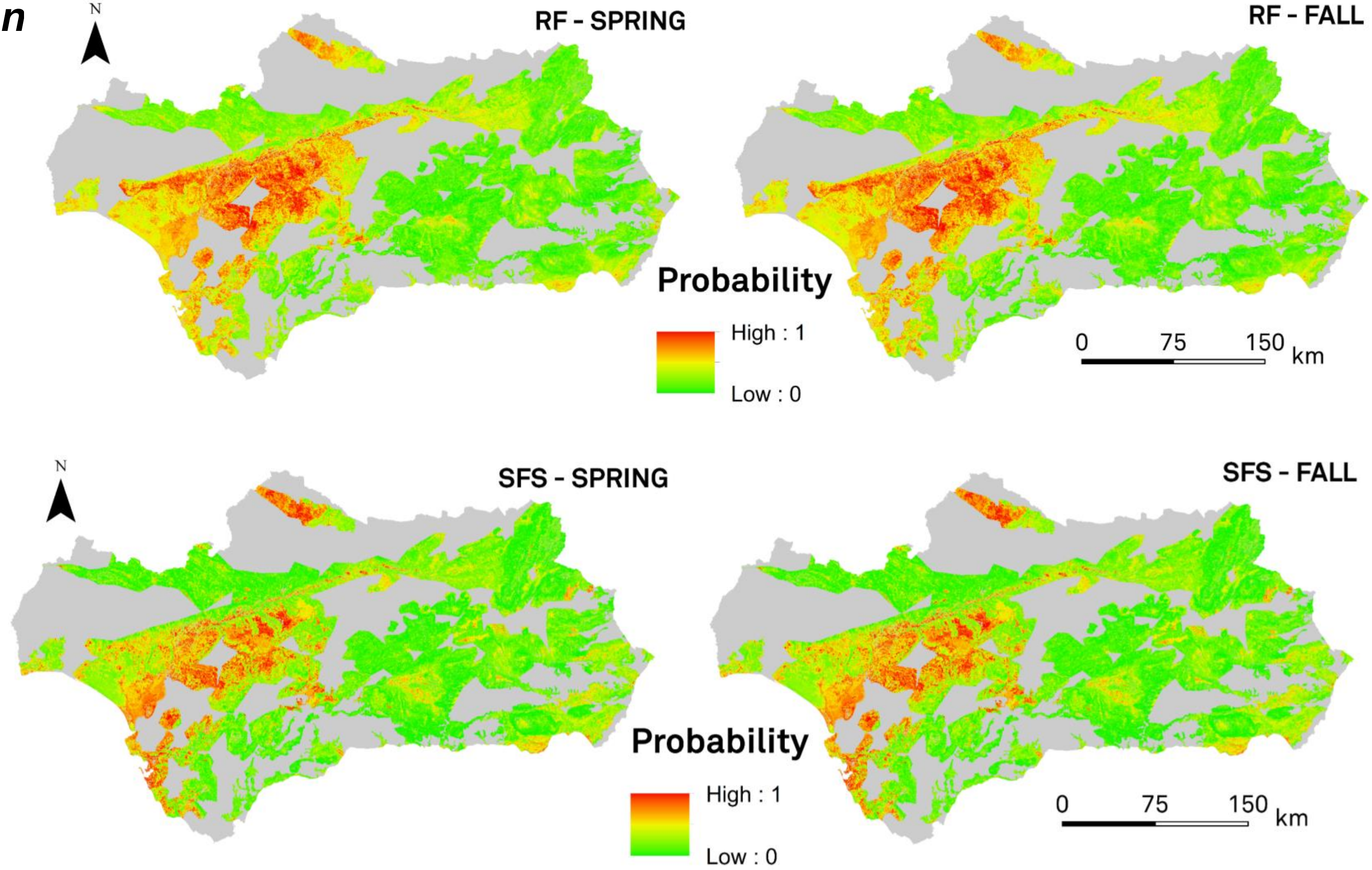
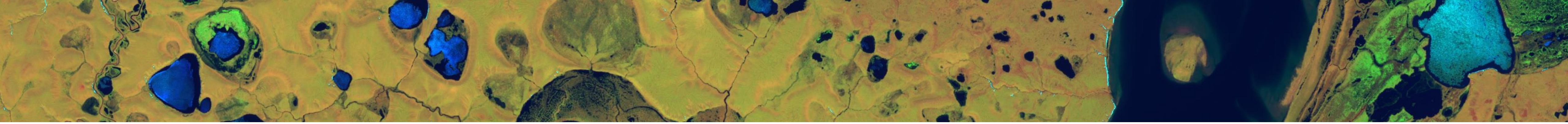


Figure 5. Prediction for 2010 of the probability of exceeding 50 mg/l of nitrates in groundwater by RF and SFS .

cc BY



4. Results & discussion

The analysis of the validation measures showed that both models had a similar error rate of 0.11 MMCE. However, the model trained by the SFS method had a similar performance with only 6 features, greatly improving the model's interpretation capacity. Some additional metrics, such as Kappa index and producer and user accuracy were obtained, generating similar results.

Several conclusions were obtained from this research. (1) First, the successful results obtained in predicting nitrates in groundwater using only features extrinsic to the groundwater bodies, which is useful for the management of large administrative regions and allows for the establishment of concrete action plans. (2) Second, the use of phenology has incorporated dynamic vegetation properties, which are key to the study of agricultural productivity, extension and timing. (3) Finally, the possibility of making a prediction in the present time, with a model that learns from past time situations. The use of feature space reduction methods such as FS has allowed to narrow down the main explanatory factors without reducing accuracy.

Table 1. Metrics for assessing the accuracy of spatial predictions.

Method	**Kappa**	**MMCE**	**Overall accuracy**	**Producer**	**User**
RF	0.73	0.11	89.63%	0.85	0.92
SFS	0.74	0.11	89%	0.86	0.90

HS8.2.2. – Groundwater in a changing environment. Sustainability and adaptative management of resources.

Thank you!

EGU General Assembly 2020

Background image source: Sentinel. Flickr.