Tuning spatial parameters of Geographical Random Forest: the case of agricultural drought

Daniel Bicák*

Department of Applied Geoinformatics and Cartography, Faculty of Science, Charles University, Czechia
* Corresponding author: daniel.bicak@natur.cuni.cz

ABSTRACT
Machine learning algorithms are widely used methods in geographical research. However, these algorithms are not properly exploiting the underlying spatial relationships present in the geographical data. One of the approaches, which addresses this problem, is based on an ensemble of local models, which are constructed from samples in close proximity to the location of prediction. This concept was applied to the Random Forest (RF) algorithm, creating a Geographical Random Forest (GRF). This study aims to further develop GRF by tuning the spatial parameters for each location in case of agricultural drought. In addition to tuning, the explanatory property of RF within the framework GRF is explored. Four machine learning models were constructed; regular RF, regular RF with spatial covariates, GRF, and GRF with the tuning of spatial parameters. Models were evaluated using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Although the decrease in RMSE in this very case is relatively small, the method may provide higher improvement with different datasets.

KEYWORDS
machine learning; Random Forest; Geographical Random Forest

Received: 9 December 2023
Accepted: 29 September 2023
Published online: 1 November 2023

Bicák, D. (2023): Tuning spatial parameters of Geographical Random Forest: the case of agricultural drought.
AUC Geographica 58(2), 187–199
https://doi.org/10.14712/23361980.2023.14

1. Introduction

Machine learning algorithms are increasingly being used in various academic and commercial research fields. Geographical topics are no exception. The difference between machine learning in geography and other research fields is the input data. Geographical data are located in space, therefore are often denoted as spatial data. Location in space is described either in absolute terms (geographical coordinates) or in relative, for example, adjacency to neighborhood entities. This property of data can be exploited to achieve a higher degree of accuracy. Such an approach can be described as spatially sensitive. This study aims to apply a spatially sensitive machine learning model to the complex phenomena of agricultural drought.

The traditional approach, which does not take into account spatial patterns inside data, suffers from the inability to properly model spatial relationships. This inability is called "spatial non-stationarity" (Fotheringham, Charlton and Brunsdon 1996, 605). Fotheringham, Brunsdon and Charlton (2003, 9–10) list three reasons which cause spatial non-stationarity. Firstly, there is sampling variation, which relates to statistical artifacts. Secondly, some relationships are intrinsically different across space, especially for social processes. And lastly, there is a possibility that one or more important variables are missing from the model. The second point can be exemplified by Simpson's Paradox (Simpson 1951) which refers to incorrect estimation of function when data are analyzed separately and then aggregated.

One option to capture spatial non-stationarity is to include spatial covariates, the most popular and easiest to use being geographical coordinates. A comprehensive study and evaluation of such an approach were conducted (Hengl et al. 2018) with Random Forest (RF) algorithm. The second option is to create an ensemble of local models, which encompass only a portion of all samples depending on their location. For each location where prediction takes place, a local model is created including n closest samples. In addition, one global model is created and final predictions are weighted averages of global and local models. This method has been applied to Linear Regression creating Geographical Weighted Regression (Brunsdon, Fotheringham and Charlton 1996) and to the RF algorithm creating Geographical Random Forest (GRF) (Georganos et al. 2019).

1.1 Random Forest

The Random Forest was developed by Breiman (2001) and belongs to a family of decision trees. Decision tree-based models make predictions by dividing prediction space into several subregions and have a tree-like hierarchical structure. The building of decision trees follows two steps; firstly, at each split, divide the feature space (range of values for each feature) into several distinct regions. Secondly, for each observation that falls into the same region, a prediction is made – the mean of values of the predicted variable. The problem is to find value by dividing predictor space most efficiently. The threshold value is calculated so that the overall sum of square errors is minimized (Kuhn and Johnson 2013, 175). However, it is not computationally feasible to find optimal partition for features. The algorithm begins with one region and then successively divides the feature space at each split. At each split, the best partition is made at that particular split. This is also known as a top-down greedy approach. Unfortunately, the variance of trees is very high. The application of bootstrap aggregation (bagging) decreases the variance by averaging many similar trees from bootstrapped datasets. The new datasets are sampled with replacements from the original dataset. On average, one-third of all samples are not used during the tree-building process and are called out-of-bag (OOB). RF algorithms further develop this concept by incorporating an ensemble of decision trees. In addition, the algorithm considers only part of the features at every splitting, which decreases correlation among trees and therefore decreases variance.

The Random Forest algorithm achieves one of the highest forecasting accuracies compared to other algorithms for the broad field of tasks (Berk 2020, 288). One of the advantages of RF is its great performance for high dimensional data when the amount of predictors is higher than the amount of observation. Another reason to choose RF is its great computational performance, which is native to all tree-based algorithms. "Compared to bagging, RF is more computationally efficient on a tree-by-tree basis since the tree-building process only needs to evaluate a fraction of the original predictors at each split" (Kuhn and Johnson 2013, 200). RF can be running simultaneously on more cores and results can be aggregated afterward (Liaw and Wiener 2002, 22).

Hyperparameters of machine learning algorithms control the training process. Values for each hyperparameter need to be set before the start of the training phase. Hyperparameter optimization is necessary to construct a stable and accurate machine learning model. RF has several important hyperparameters. A number of randomly drawn features during the splitting phase often denoted as mtry, influences the stability and prediction accuracy. Lower values tend to boost the stability of the model, on the other hand, the accuracy is slightly lower (Probst, Wright and Boulesteix 2018, 3). Lower values also decrease the computational complexity, as the algorithm does not need to calculate as many thresholds. The next parameter, the number of trees in the forest should be set to at least 100. According to Probst and Boulesteix (2017), the accuracy increases with diminishing returns when inputting higher values. However, the computational complexity increases as well.

The Hyperparameter Minimal Number of Samples describes how many samples are used for training each tree and its effect is similar to hyperparameter number of randomly drawn features. Additional parameters are Node size (minimum number of samples in a terminal node) and Splitting rule (function to assess the quality of the split). Generally, RF performs well with tuning only mtry hyperparameter (Fernández-Delgado et al. 2014, 3175). Hyperparameters can be tuned with traditional methods, for example, Grid Search or Random search. Existing OOB samples can be utilized to evaluate the model, which saves time.

In addition to hyperparameters native to RF, GRF brings hyperparameters bandwidth and local weight. Bandwidth describes the size of the kernel for each local model. In other words, a number of closest samples of which local models are trained. There are two types of kernels – adaptive and fixed. The former encompasses n closest samples in the vicinity where prediction takes place. The latter is the circle, in which the radius is the bandwidth (Fotheringham, Brunsdon and Charlton 2003, 44). The final prediction for location is made from a weighted average of the local model and global model, where the weight for the local model is a tunable parameter. The combination of two models results in higher accuracy – the local model ensures low bias and the global one has low variance (Georganos et al. 2019, 7). The drawback is higher computational complexity, GRF needs to compute a new model for each predicted location.

1.2 Agricultural drought

Environmental hazards are natural phenomena that negatively affect human society regarding economic and social losses. Drought hazard belongs to the most damaging and widespread causes of huge economic and human losses. The severity of drought depends on the environment's (or society's) ability to cope with hazards. For example, in developed countries, drought's direct impact is almost invisible, and indirect impact projects to higher consumption of water to irrigate agricultural plants. In developing countries, drought might cause crop failure and subsequent instability. However, climate change will worsen many aspects of drought, including its recurrence, severity, and timespan (Mukherjee et al. 2018).

The gravity of drought hazard is reflected in the abundance of research studies focusing on identifying vulnerable locations or factors. Various methods have been applied; including the subjective weighting of drought drivers (Wilhelmi and Wilhite 2002) or the analytical hierarchy process (Hoque et al. 2020). More recently, machine learning algorithms are utilized to model drought. For example, Rahmati et al. (2019) employed RF, Support vector machines, and others to create a vulnerability map of Queensland, Australia. A similar study utilized an Artificial Neural Network (Rahmati et al. 2020).

The machine learning model requires independent variables, which influence the drought intensity, and a dependent variable, which functions as a drought indicator. According to Mishra and Singh (2010, 207), the drought indicator is a prime variable for assessing the effect of drought and defining different drought parameters, which include intensity, duration, severity, and spatial extent. The selection of an appropriate indicator is essential as it will be the dependent variable, which will be predicted by the model. Soil moisture, especially within the root system of the plants, is an accurate indicator of agricultural drought. Soil moisture-based indicators are used in similar studies concerning agricultural drought e.g. Rahmati et al. (2019) and Rahmati et al. (2020).

The severity of agricultural drought is influenced by various factors. The most profound is meteorological. The connection between agricultural drought and meteorological patterns is clear. Precipitation is the only source of moisture for the environment with the exception of irrigation, which is available for a fraction of cultivated areas. Temperature influences the rate of transpiration, higher temperatures increase the transpiration rate. A region with higher temperatures is, therefore, more prone to drought. However, precipitation deficit impacts are greater than high temperatures in general (Yang et al. 2020, 9).

Topographic characteristics refer to the quantitative descriptions of the physical features of land. Vegetation in mountainous regions subscribes to different patterns of climatic conditions and develops specific adaptations. The slope of an area affects the run-off, recharge, and movement of surface water. Flat terrain areas have relatively high infiltration rates, on the other hand, areas with steeper slopes have low infiltration rates and higher run-off (Shekhar and Pandey 2015, 409). Another topographic factor is aspect, which refers to the orientation of the slope. The aspect of a slope can influence local climate because of the length of the exposure to sun rays. West and south-facing slopes are warmer than east and north-facing slopes, therefore having lower soil moisture and higher evaporation rate (Magesh and Chandrasekar 2010, 375). The topographical Wetness Index (TWI) (Beven and Kirkby 1979) describes the proclivity of a place to accumulate water based on topographic information. TWI is a widely used indicator to obtain information on the spatial distribution of wetness conditions, since only a terrain model is required for calculation. Soil properties are important factors influencing the environment's ability to cope with drought. Soil acts as a substrate for plants' roots, providing them with water and nutrients. Soil characteristics influence these functions to various degrees. Soil texture refers to the size of solid particles, that soil is composed of. The size of particles determines the amount of water that can be stored for plants. Organic matter is one of the most important soil characteristics. According to Bot and Benites

(2005, 35–36), organic content increases water infiltration and water holding capacities, increasing the diversity and activity of soil organisms and providing nutrient availability. Land cover is intertwined with water demand and the coping abilities of the environment to drought hazards. Land use describes how society uses land, land cover refers to the physical features of the land. In case of vulnerability to drought, the scientific community classifies several types – mainly agricultural fields, grasslands, forests, barren lands, urban areas, and water bodies (Jain, Pandey and Jain 2014; Thomas et al. 2016; Hoque et al. 2020).

This study continues to develop GRF by tuning spatial parameters for each location. The hypothesis is that tuning the spatial hyperparameters for each location will improve the accuracy of the GRF model. This hypothesis is based on the assumption that those spatial hyperparameters are spatially correlated. The study aims to confirm the hypothesis by completing three subtasks; firstly, creating an accurate statistical model based on the RF algorithm of drought hazard which consists of many local models and one global model and subsequently evaluating the accuracy metrics for both models. Secondly, performing a tuning of parameters for each local GRF model, answering the question of whether it is possible to improve the accuracy of the model further. Lastly, providing insight into the vulnerability modeling from the feature's importance of local models.

2. Methodology and Data

This section describes the study area, datasets used to build the machine learning model, and method, which facilitates the local tuning of spatial hyperparameters.

2.1 Study Area

The problem is studied within the agricultural landscape in the Czech Republic, Central Europe. The extent reaches approximately 50 km beyond the border north to Poland, west to Germany, south to Austria, and east to Slovakia. The study area was limited to an agricultural landscape with these conditions:

- Forest should not cover more than 20% of the pixel area.
- Built-up areas should not cover more than 20% of the pixel area.

The study are is shown in Fig. 1. The total area of locations, that met the conditions is 53,860 km^2.

2.2 Data

The drought model requires many features (independent variables) to work properly. This is reflected in the variety of data sources, from which data were collected. The independent variable – the drought predictor was chosen to be the Soil Water Index (SWI). The selection was influenced by the availability of data in terms of spatial and temporal resolution. SWI is available within Copernicus Global Land Service (Bauer-Marschallinger et al. 2018). SCATSAR-SWI (Scatterometer Synthetic Aperture Radar Soil Water Index) is computed from the data fusion of products Sentinel-1 SSM (Surface Soil Moisture) and ASCAT SSM/SWI, which assess soil moisture. The dataset includes layers with various temporal parameters T, which correspond to different soil depths. The layer with a T value of 20 was chosen as it correlates best with the subsoil conditions (10–20 cm below the surface) (Paulik et al. 2014, 5) and has uniform quality scores across the study area.

Meteorological features were acquired from the E-OBS dataset maintained by the European Climate

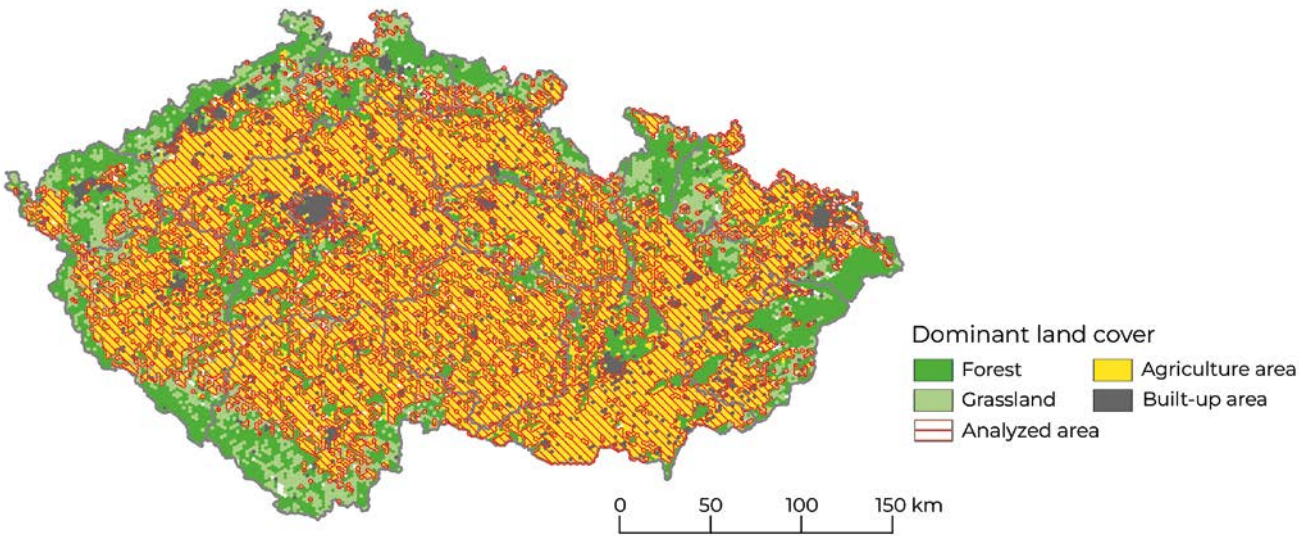

Fig. 1 Study area.

Tab. 1 Aggregated thematic classes and their corresponding Land cover categories.

Aggregated class	Former classes
Built-up areas	Continuous urban fabric, Discontinuous urban fabric, Industrial or commercial units, Road and rail networks and associated land, Port areas, Airports, Mineral extraction sites, Dump sites, Construction sites
Agricultural areas	Non-irrigated arable land, Vineyards, Fruit trees and berry plantations, Annual crops associated with permanent crops, Complex cultivation patterns
Grasslands	Pastures, Natural grasslands, Moors and heathland
Forests	Broad-leaved forest, Coniferous forest, Mixed forest, Transitional woodland-shrub

Assessment & Dataset project. E-OBS is interpolated from point data gathered from national meteorological stations across Europe. According to the project website (Cornes et al. 2018), Czechia has an above-average density of stations (770 km^2 for precipitation and 913 km^2 for temperature per station). All topographic-related features were extracted from European Digital Elevation Model (EU-DEM).

All soil properties except Soil organic matter were acquired from the "Topsoil Physical Properties for the Europe" dataset, which is based on Land Use and Cover Area frame Statistical Survey (LUCAS) dataset. LUCAS is the largest harmonized soil dataset in Europe overseen by the Statistical Office of the European Union, which consisted of in situ measurements from more than 22,000 locations (Orgiazzi et al. 2018). Another dataset derived from LUCAS is Soil Organic Matter (SOM) fractions (Cotrufo et al. 2019), which utilized more than 9400 points, to interpolate point data to a grid with a 1 km spatial resolution using the RF algorithm. Organic matter is divided by size into particulate and mineral-associated organic matter (less than 53 µm). Datasets are delivered in GeoTiff format and ETRS89-LAEA coordinate system. Both datasets, TPPE and SOM are distributed by the European Soil Data Centre (Panagos et al. 2012).

Land cover information was obtained from Corine Land Cover (European Environment Agency (EEA), 2019). Land cover categories were aggregated into four thematic classes – built-up areas, agricultural areas, grasslands, and forests. Land cover categories and their corresponding thematic classes are listed in the table below (Tab. 1).

Metadata of datasets used in the study are listed in following table (Tab. 2).

2.3 Methods

A new variant of GRF was developed – Locally Tuned GRF (LT GRF). Values for hyperparameters bandwidth and local weight are universal for every sample across space. LT GRF aims to find optimal values for each location. The optimal values are found for each location during the training process. Values are interpolated for the whole study area by linear interpolation. In case there are several different values in one place, the mean value is used. We assume, that there exists a spatial autocorrelation in the model parameter's weights and bandwidth. Six models in total were developed; RF, GRF, and LT GRF with coordinates and RF, GRF, and LT GRF without coordinates. LT GRF should be the most accurate model created. Algorithm LT GRF can be described by the following pseudocode.

Algorithm 1 Geographical Random Forest with local tuning.

1: **for** training observation **do**
2: **for** bandwidth, local weight in the kernel, weights **do**
3: Perform Random Forest.
4: Perform Random Forest with bandwidth number of samples.
5: **end for**
6: Select optimal values of bandwidth and local weight.
7: **end for**

Tab. 2 Metadata of datasets.

Product name	Original temporal resolution	Temporal resolution used in the study	Original spatial resolution	Format	Reference
SCATSAR-SWI	1 day	14 days	1 km	NetCDF	Copernicus Global Land service (2023)
E-OBS	1 day	14 days	0.1°	NetCDF	Cornes et al. (2018)
EU-DEM	/	/	25 m	GeoTiff	European Environment Agency (2016)
Topsoil Physical Properties for the Europe	/	/	500 m	GeoTiff	Panagos et al. (2022)
Soil Organic Matter	/	/	1 km	GeoTiff	Lugato et al. (2021)
Corine Land Cover	/	/	100 m	GeoTiff	Copernicus Land Monitoring Service (2019)

8: Interpolate bandwidth and local weight values for the location of testing observations.
9: **for** testing observation **do**
10: Perform Random Forest.
11: Perform Geographical Random Forest.
12: Compute the weighted average of the output of Random Forest and Geographical Random Forest.
13: **end for**

The process of evaluating models consists of several steps; pre-processing of the data, model building, and performance evaluation by accuracy metrics. Firstly, the time periods were chosen for SWI. Five time periods – the first two weeks of August from 2015 to 2019. Each period has a different distribution of values, therefore combined dataset contains observations with low and high values for the same place.

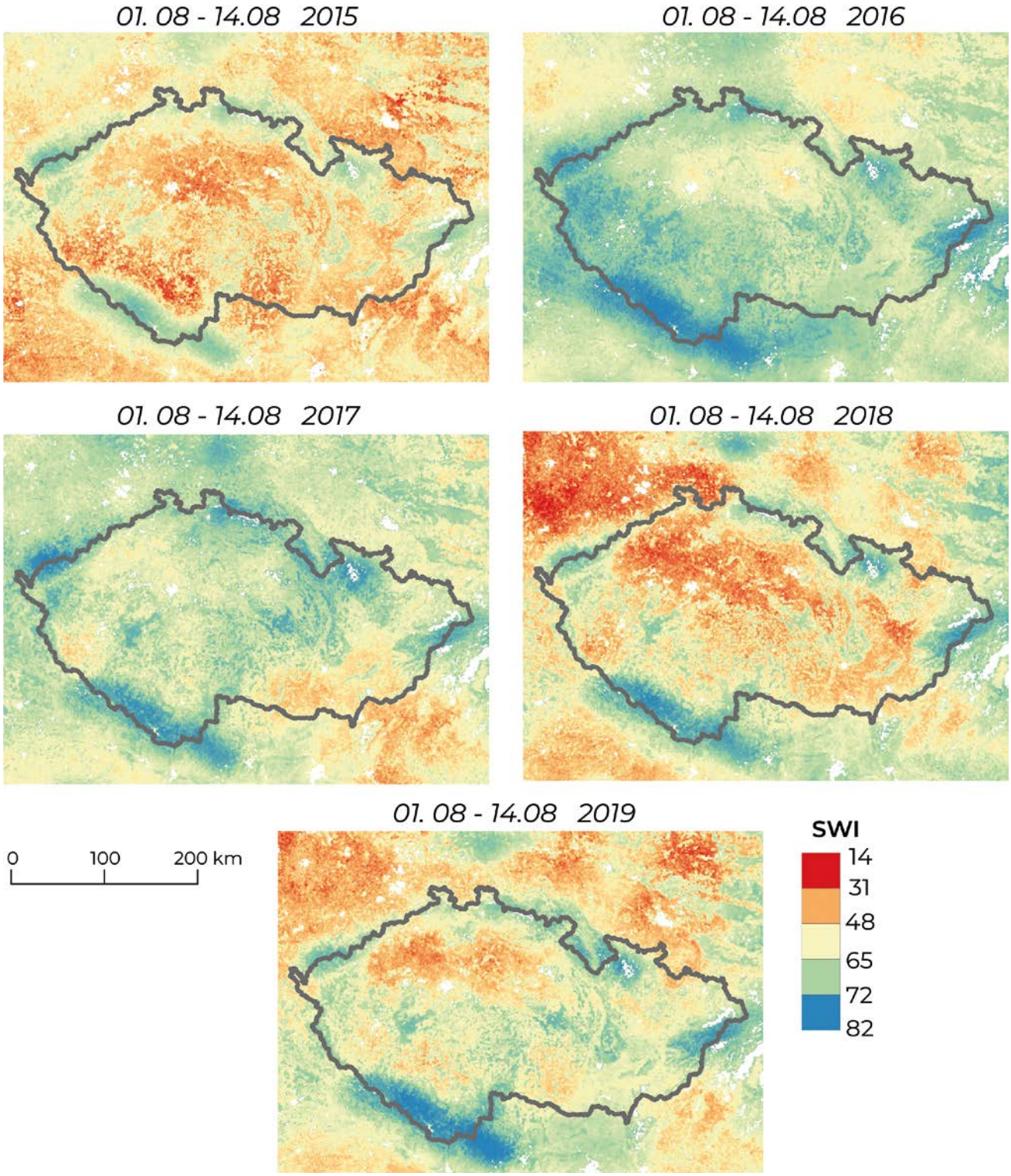

Fig. 2 SWI values for each selected period.

Various values of SWI for the same place can facilitate a robust and accurate model. SWI values are shown in Fig. 2. Secondly, drought predictors were chosen; geographical coordinates (X and Y), elevation, slope, aspect, TWI from category terrain characteristics, temperature and precipitation from meteorological characteristics, soil texture, organic matter content, soil bulk ratio, and AWC from soil properties. Organic matter content was created as a sum of both layers of the SOM dataset. Landcover features are represented by their proportion in each location. Four land cover classes were chosen – built-up areas, agricultural areas, grasslands, and forests. In addition to the listed features, the distance to large water bodies (rivers and reservoirs) was added. SWI and meteorological features are available for each day, therefore need to be aggregated. SWI and temperature are averaged, and precipitation is summed. Three periods of temperature and precipitation are selected. Two weeks period, which is identical to the SWI period, a one-month period (two weeks before the start of the SWI period), and a three-month period.

All datasets were resampled to sample size with a resolution of 3 km^2 using linear interpolation. Values of all features (independent variables) were scaled from 0 to 1. The dataset was split into training and testing sets with a ratio of 0.33 (two-thirds were used for tuning and one-third for testing) using random sampling. Three hyperparameters of RF were tuned using Grid Search with OOB samples – a number of randomly drawn features, a number of trees, and a minimal number of samples. Subsequently, GRF hyperparameters bandwidth and local weight were tuned using a grid search cross-validation method. Several bandwidth values were tested; 50, 100, 150, 200, 250, 500, 750, 1000, 1500, 2000 and 5000. The distances are not equal for all locations, because of the use of an adaptive kernel. For a bandwidth of size 50, the average distance is 5205 m. For maximum bandwidth of 5000, the average distance is approximately 45 km. The parameter of local weight values from 0 to 1 with increment 0.1 were tested.

The performance of statistical models is evaluated by metrics – Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). These metrics were chosen because of their wide use in the scientific community. The RMSE is calculated by the formula:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

and MAE is expressed by formula:

$$MAE = \frac{1}{n} \sum_{i=1}^{n-1} |y_i - \hat{y}_i|$$

In addition to the RMSE and MAE, relative accuracy in % is used. Relative accuracy is calculated as a ratio of error (RMSE or MAE) to the range of values of the dependent variable without outliers. Outliers are understood to be values less than one percentile and higher than 99 percentil of SWI values.

The import, processing and building of the models took place using the python programming language. Libraries used include scikit-learn, NumPy, Pandas and Xarray. Map outputs were created using QGIS software.

3. Results

This section describes the results of tuning the machine learning models, performance assessment, and the feature importances of GRF LT model shown graphically in the maps.

The hyperparameters were tuned using the Grid Search method with OOB samples. The optimal value for the number of randomly drawn features was found to be 16. The minimum number of samples was set to 5. The number of trees is 200, and the RMSE decreases with diminishing returns with increasing the value of the hyperparameter. The RMSE of values of spatial hyperparameters is depicted in Figure 3. The optimal value for local weight was found to be 0.7 (0.7 local model and 0.3 global model). The decrease in error between regular RF (local weight is 0) is approximately 0.297 RMSE. In comparison, the difference between the default value of hyperparameter Randomly Drawn Features (typically one-third of all available features) is 0.2 RMSE. The optimal bandwidth is 100 observations. The decrease in RMSE in comparison to the global model is very small 0.09, smaller than the decrease of hyperparameter local weight.

The optimal values of the bandwidth and local weight are found for each location. The count of each value of the parameter is displayed in a histogram below (Fig. 4). The most numerous value for bandwidth is 50 constituting 20% of all values. The second place belongs to value 100 with a 12.7% share. Other values constitute a portion smaller than 10%. GRF assigns one universal value to all locations, however, as can be seen, it is not optimal for the vast majority of locations. In the case of local weight, the situation is more uneven. The most numerous local weight value is 1 (only the local model is employed) which constitutes 56.4% of all values. The second most numerous is value 0 (only the global model is employed) with an 8% contribution to all values. Other values are represented less, the count decreases with lower local weight. However, the best value achieved by GRF tuning is 0.8. This value is not optimal for more than 92% of all locations.

Each tuned model was trained and tested. RF without coordinates achieved an RMSE of 4.48 (MAE of

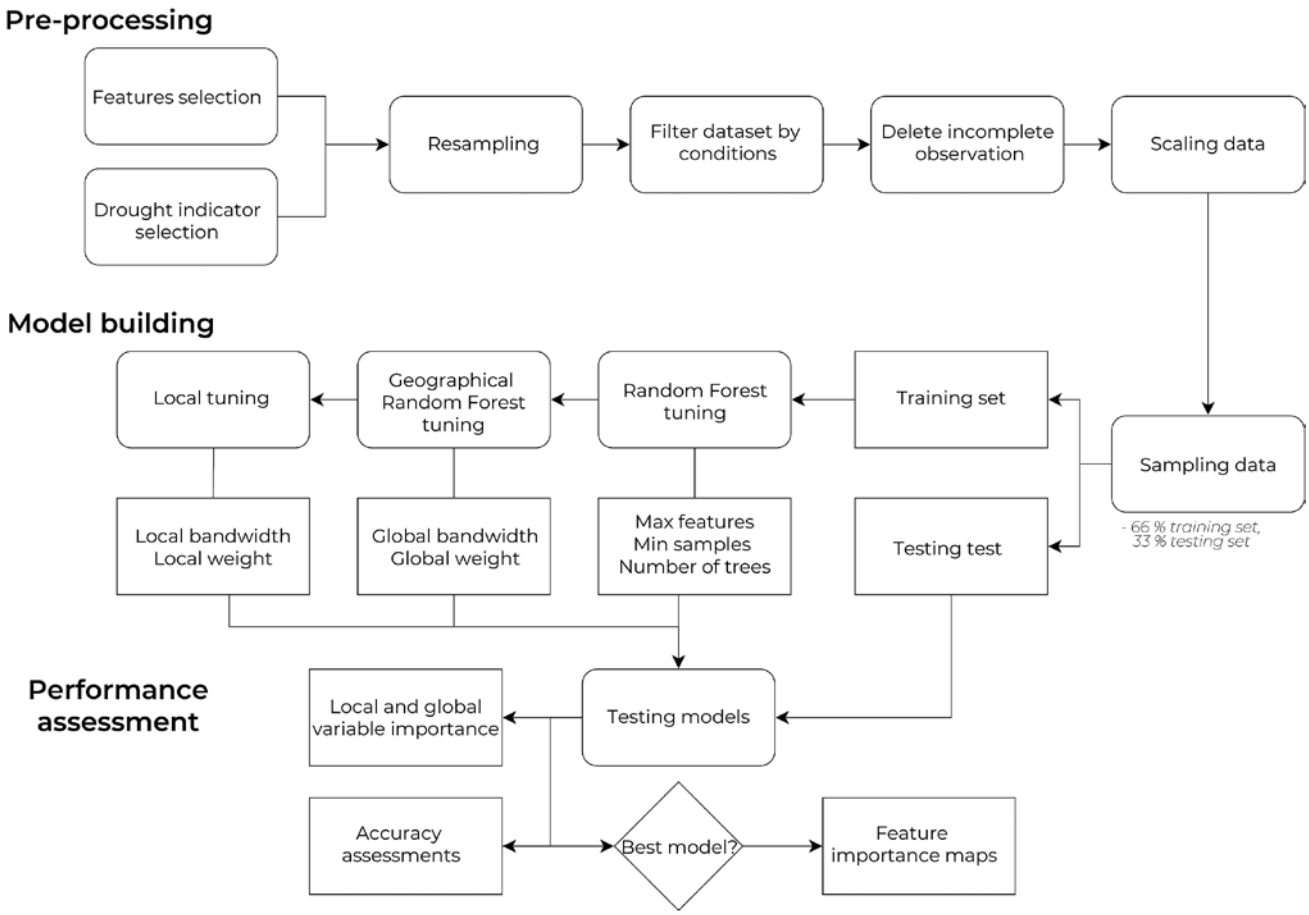

Fig. 3 Flow chart of the research.

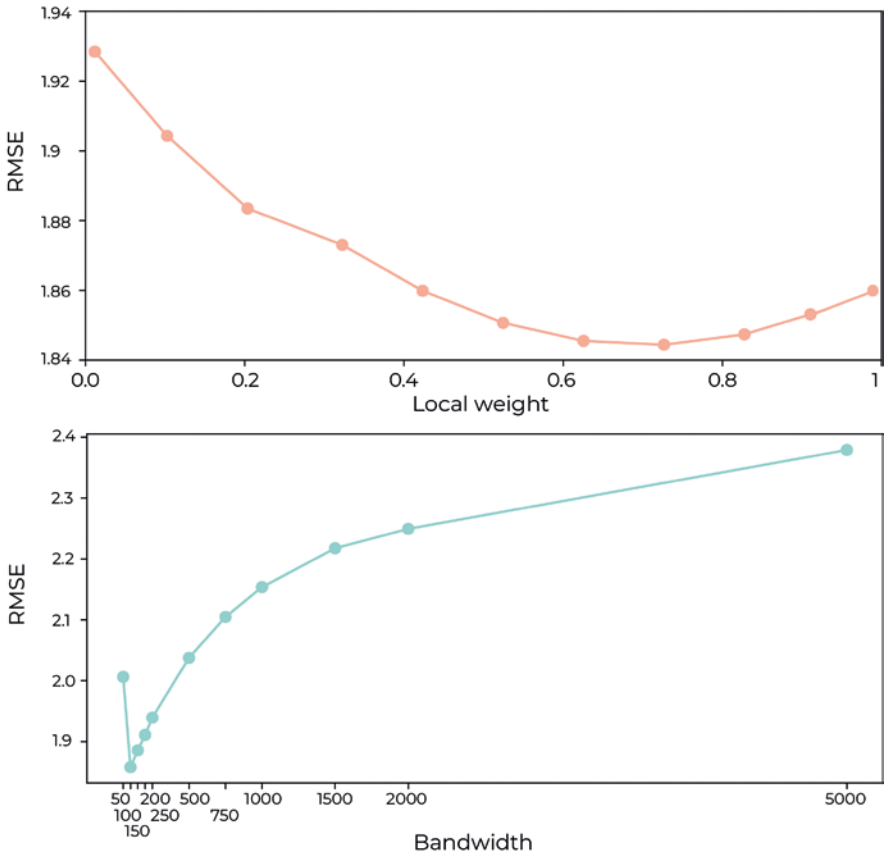

Fig. 4 RMSE of spatial parameters, extracted from tuning of GRF with spatial coordinates.

Tab. 3 Accuracy metrics for each tested model expressed in relative and absolute values. Models with spatial coordinates are denoted with 'XY'.

	RMSE		MAE	
	abs	rel [%]	abs	rel [%]
RF model	4.48	90.86	3.40	93.02
RF XY model	2.70	94.46	2.05	95.77
GRF model	2.60	94.66	1.96	95.97
GRF XY model	2.42	95.03	1.83	96.24
LT GRF model	2.58	94.70	2.02	95.85
LT GRF XY model	2.41	95.05	1.80	96.30

3.4), RF with spatial coordinates achieved an RMSE of 2.7 (MAE of 2.05), GRF without coordinates of 2.6 (MAE of 1.96), GRF with spatial coordinates of 2.42 (MAE of 1.83), LT GRF without coordinates of 2.58 (MAE of 2.02) and LT GRF with spatial coordinates of 2.41 (MAE of 2.41). Values are listed in the table below (Tab. 3).

Feature importance of the RF and GRF models are displayed in bar plots below (Fig. 5 and Fig. 6). The most important features of the RF (without spatial coordinates) are precipitation features. Together they account together for almost approximately 60% of the importance. Summed precipitation over 1-month accounts for 38% and is the most important feature. The fourth place belongs to elevation with 6% of importance. The elevation is followed by temperature features, each accounting for 5% of importance. Other features have less than 3% of importance.

Feature importances of the GRF LT model (without spatial coordinates) were aggregated by the mean value (Fig. 7). Values are similar to the RF model. The

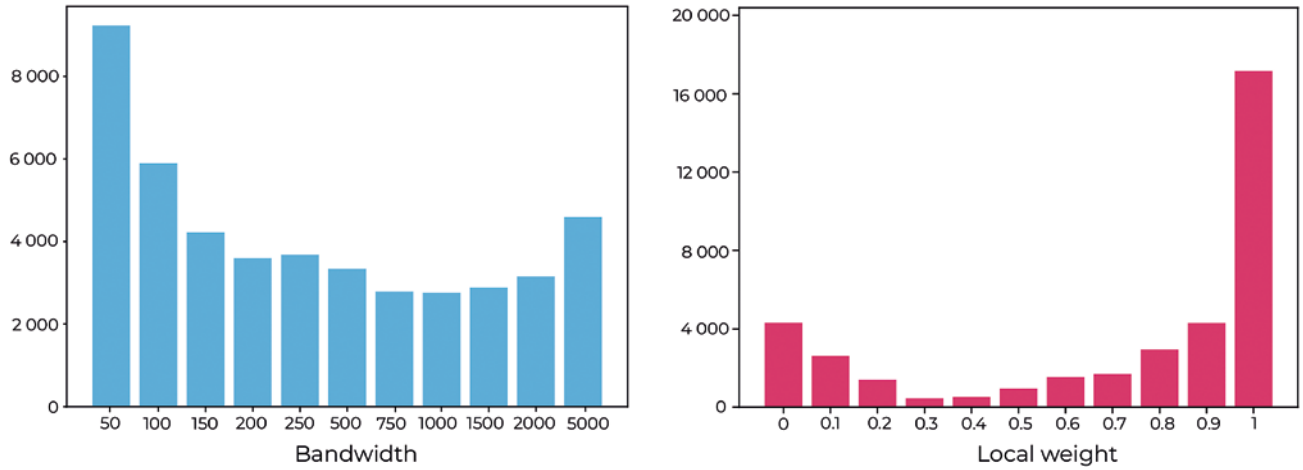

Fig. 5 Histograms for bandwidth and local weight extracted from tuning of GRF with spatial coordinates.

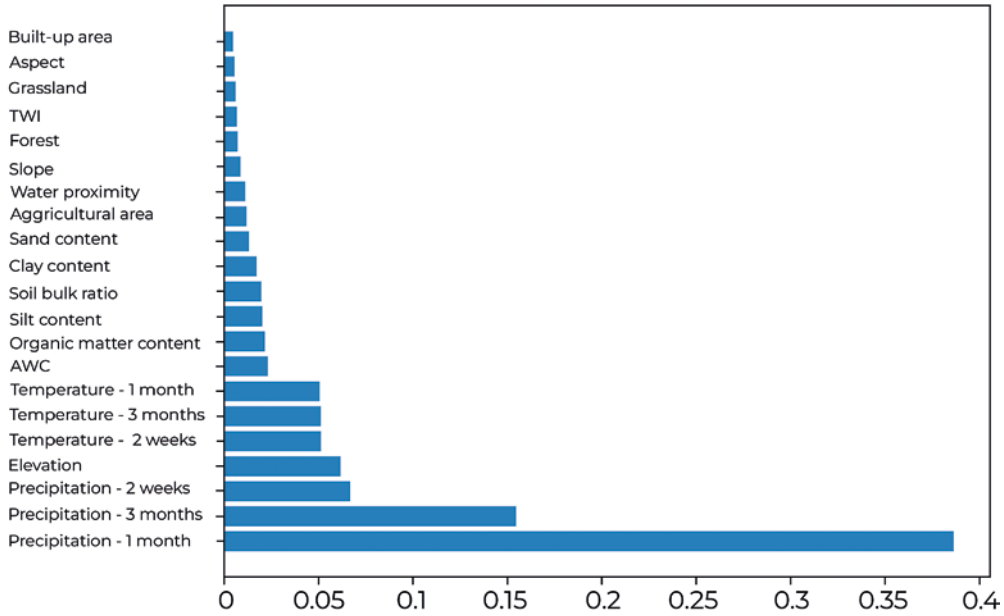

Fig. 6 Feature importances for RF model (without spatial coordinates).

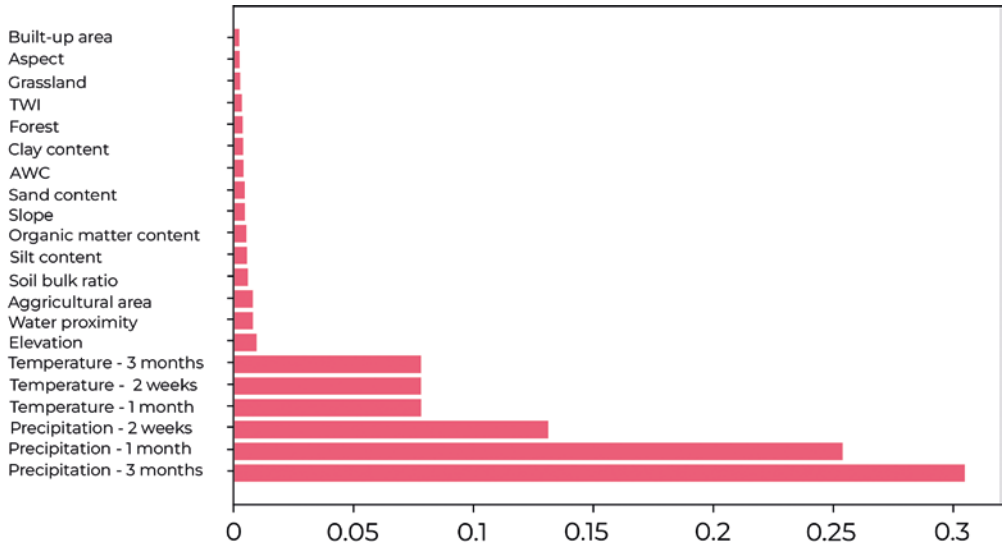

Fig. 7 Feature importances for GRF LT without spatial coordinates.

most important features are precipitation features (1 month with 38%, 3 months with 15%, and 2 weeks with 6%) followed by temperature features (each with approximately 7%). Elevation and water proximity account for approximately 1%. The rest of the features account for only 6% of feature importance.

The top eight features were visualized in maps (Fig. 8). Precipitation features show different spatial distributions. The longest precipitation period (3-months) shows strong importance in the northwestern part of the area, 1-month precipitation in western and southwestern parts of the area, and the shortest period in the eastern part. Temperature features show a similar pattern for each period. Strong importance can be assessed in southern part of the area.

3. Discussion

The tuning of spatial parameters showed that there is no optimal uniform value for all training locations. However, GRF or LT GRF did not achieve a significantly higher degree of accuracy. The minuscule difference in error between models can be explained in several ways. Firstly, regular RF achieves very good results. Accuracy of more than 90.86% is very high and the possibility for improvements is limited. It is most likely that model accuracy cannot be significantly improved any further for the given modeled problem and available input data sets. Secondly, spatial non-linearity is explained well by spatial coordinates, which are input features in the RF model. In other words, the global model (RF model with spatial coordinates) has not left any space for local models (GRF and GRF LT models) to improve more significantly.

The reduction in RMSE error between the GRF and GRF LT models (models with spatial coordinates) is surprisingly low. During the tuning phase, GRF LT achieved an RMSE of 1.8, which is significantly less than the resulting error in the testing phase. The limited improvement of the GRF corresponds with the visual examination of bandwidth. There is no or very little spatial correlation between bandwidth and local weight and a decrease in error. Values are localized randomly as a residue of random error.

GRF creates local models on a subset of original datasets. This process can be reinterpreted as a huge number of created decision trees with a very small number of observations. A similar situation can be recreated with regular RF with parameter maximum samples set to a value of best bandwidth (100). However, experiments show that such a model is very inaccurate (RMSE of 9.5) and this hypothesis can be rejected.

The performance of GRF and LT GRF was compared to the performance in other studies. The GRF with spatial covariates in the study by Georganos et al. (2021) achieved an RMSE of 0.606, the global model achieved an RMSE of 0.65. The error decreased by 6.76%. Master thesis by Hokstad and Tiganj (2020) compared RF with spatial covariates to GRF. RF achieved an RMSE of 17,944 and GRF of 16,705, a 6.9% decrease in error. In both studies, a decrease in error between RF with spatial coordinates and without them is more significant.

Improvement of GRF or LT GRF over regular RF is small and computational runtime is much higher. A desktop PC is not sufficient for larger datasets (more than 100,000 samples) and a more expensive solution needs to be employed. Therefore, it may seem that GRF might not be advantageous over classical RF based on this case study given the computational requirements and not significant improvements in model performance.

The feature importance results provide unique insight into the drivers of agricultural drought.

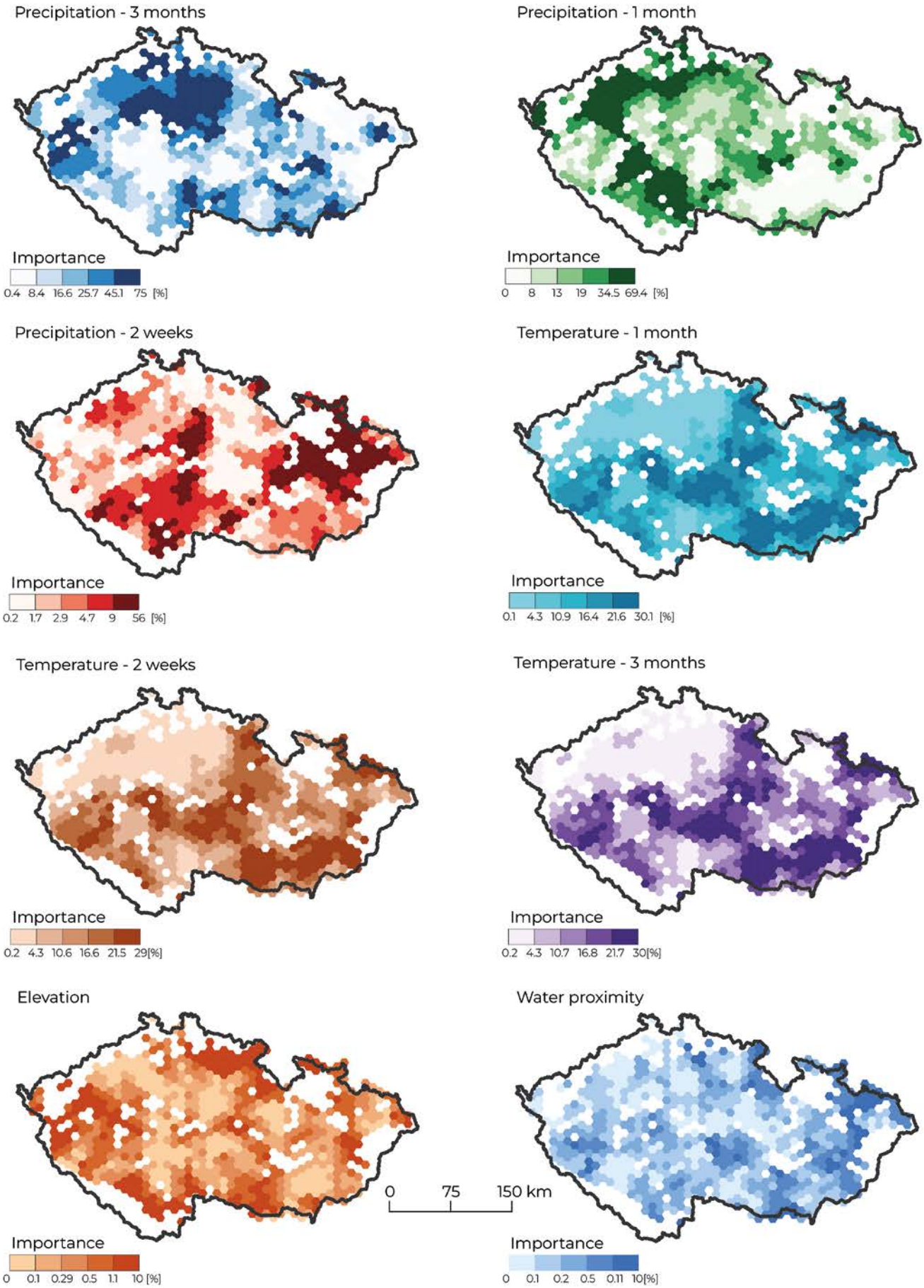

Fig. 8 Top most important features of GRF LT (without spatial coordinates) visualized on the map.

However, to draw conclusions from maps, deeper knowledge of local conditions is needed.

The concept of local sub-models and their parameters can be studied further. Despite the increase in accuracy being small, local tuning might deliver more beneficial results in different use cases. Therefore, it would be useful to find and evaluate spatial patterns in various datasets, which would benefit most from this method. The use of more sophisticated spatial interpolation methods such as kriging when obtaining unknown values of local parameters can increase the accuracy of models. Such an approach would be particularly advantageous for a sparse dataset. The GRF and GRF LT models used use a binary kernel – records up to a certain distance are included in the local model. Another approach is to use a function that would assign weights to records based on their distance.

The explanatory function of the model (features importance) has the potential to provide additional insight into geographical phenomena. Results from GRF or LT GRF can be compared with more established methods such as Geographic Weighted Regression. This concept can be also extended to other Machine Learning algorithms. As mentioned by Georganos et al. (2019), Support Vector Machines are suitable methods, because of their lower computational complexity.

4. Conclusion

The study developed six machine learning algorithms; RF with and without spatial coordinates, GRF with and without spatial coordinates and LT GRF with and without spatial coordinates. LT GRF in contrast to GRF tunes the local parameters – bandwidth and local weight for each location. The models were applied and evaluated in the case of agricultural drought. A total of 21 features were used to predict drought using a soil moisture-based index as the dependent variable. In addition, the study provides insight into the feature importance property of GRF. The increase in accuracy is relatively small in this very case, however, different datasets may provide more desirable results.

References

Bauer-Marschallinger, B., Paulik, C., Hochstöger, S., Mistelbauer, T., Modanesi, S., Ciabatta, L., Massari, C., Brocca, L., Wagner, W. (2018): Soil moisture from fusion of scatterometer and SAR: Closing the scale gap with temporal filtering. Remote Sensing 10(7), 1030, https://doi.org/10.3390/rs10071030.

Berk, R. A. (2008): Statistical Learning as a Regression Problem. In: Statistical Learning from a Regression Perspective. Springer Series in Statistics. Springer, New York, NY, https://doi.org/10.1007/978-0-387-77501-2_1.

Beven, K. J., Kirkby, M. J. (1979): A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. Hydrological Sciences Journal 24(1), 43–69, https://doi.org/10.1080/02626667909491834.

Bot, A., Benites, J. (2005): The importance of soil organic matter: Key to drought-resistant soil and sustained food production. In: Food & Agriculture Org. Breiman, L. (2001). Random forests. Machine Learning 45, 5–32, https://doi.org/10.1023/A:1010933404324.

Brunsdon, C., Fotheringham, A. S., Charlton, M. (1998): Geographically weighted regression. Journal of the Royal Statistical Society: Series D (The Statistician) 47(3), 431–443, https://doi.org/10.1111/1467-9884.00145.

Copernicus Global Land service (2023): Soil Water Index. https://land.copernicus.eu/global/products/swi.

Cornes, R. C., van der Schrier, G., van den Besselaar, E. J., Jones, P. D. (2018): An ensemble version of the E-OBS temperature and precipitation data sets. Journal of Geophysical Research: Atmospheres 123(17), 9391–9409, https://doi.org/10.1029/2017JD028200.

Cotrufo, M. F., Ranalli, M. G., Haddix, M. L., Six, J., Lugato, E. (2019): Soil carbon storage informed by particulate and mineral-associated organic matter. Nature Geoscience 12, 989–994, https://doi.org/10.1038/s41561-019-0484-6.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., Gräler, B. (2018): Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6: e5518, https://doi.org/10.7717/peerj.5518.

Hokstad, V., Tiganj, D. (2020): Spatial modelling of unconventional wells in the Niobrara Shale play: a descriptive, and a predictive approach. Master's thesis. Norwegian School of Economics.

Hoque, M. A., Pradhan, B., Ahmed, N., Sohel, Md. S. I. (2021): Agricultural drought risk assessment of Northern New South Wales, Australia using geospatial techniques. In: Science of the Total Environment 756: 143600, https://doi.org/10.1016/j.scitotenv.2020.143600.

European Union, Copernicus Land Monitoring Service (2019): Corine Land Cover. European Environment Agency (EEA), https://land.copernicus.eu/en.

European Union, Copernicus Land Monitoring Service (2016): EU – Digital Elevation Model 1.1. European Environment Agency (EEA), https://land.copernicus.eu/en.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D. (2014): Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research 15, 3133–3181, https://jmlr.org/papers/volume15/delgado14a/delgado14a.pdf.

Fotheringham, A. S., Charlton, M., Brunsdon, C. (1996): The geography of parameter space: an investigation of spatial non-stationarity. International Journal of Geographical Information Systems 10(5), 605–627, https://doi.org/10.1080/02693799608902100.

Fotheringham, A. S., Brunsdon, C., Charlton, M. (2003): Geographically weighted regression: the analysis of spatially varying relationships. John Wiley & Sons. Chichester.

Georganos, S., Grippa, T., Gadiaga, A. N., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., Kalogirou, S. (2019): Geographical random forests: a spatial

extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto International 36(2), 121–136, https://doi.org/10.1080/10106049.2019.1595177.

Jain, K. V., Pandey, R. P., Jain, M. K. (2015): Spatio-temporal assessment of vulnerability to drought. Natural Hazards 76, 443–469, https://doi.org/10.1007/s11069-014-1502-z.

Kuhn, M., Johnson, K. (2013): Applied predictive modeling. Springer. New York, https://doi.org/10.1007/978-1-4614-6849-3.

Liaw, A., Wiener, M. (2002): Classification and regression by randomForest. Race News 2, 18–22, https://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf.

Lugato, E., Lavallee, J. M., Haddix, M. L., Panagos, P., Cotrufo, M. F. (2021): Different climate sensitivity of particulate and mineral-associated soil organic matter. Nature Geoscience 14, 295–300, https://doi.org/10.1038/s41561-021-00744-x.

Magesh, N. S., Chandrasekar, N., Soundranayagam, J. P. (2011): Morphometric evaluation of Papanasam and Manimuthar watersheds, parts of Western Ghats, Tirunelveli district, Tamil Nadu, India: a GIS approach. Environmental Earth Sciences 64, 373–381, https://doi.org/10.1007/s12665-010-0860-4.

Sourav, M., Mishra, A., Trenberth, K. E. (2018): Climate Change and Drought: A Perspective on Drought Indices. Current Climate Change Reports 4, 145–163, https://doi.org/10.1007/s40641-018-0098-x.

Paulik, C., Dorigo, W., Wagner, W., Kidd, R. (2014): Validation of the ASCAT Soil Water Index using in situ data from the International Soil Moisture Network. International journal of Applied Earth Observation and Geoinformation 30, 1–8, https://doi.org/10.1016/j.jag.2014.01.007.

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, J., Fernández-Ugalde, O. (2018): LUCAS Soil, the largest expandable soil dataset for Europe: a review. European Journal of Soil Science 69(1), 140–153, https://doi.org/10.1111/ejss.12499.

Panagos, P., Liedekerke, M. V., Jones, A., Montanarella, L. (2012): European Soil Data Centre: Response to European policy support and public data requirements. Land Use Policy 29(2), 329–338, https://doi.org/10.1016/j.landusepol.2011.07.003.

Panagos, P., Liedekerke, M. V., Borrelli, P., Köninger, J., Ballabio, C., Orgiazzi, A., Lugato, E. (2022): European Soil Data Centre 2.0: Soil data and knowledge in support of the EU policies. European Journal of Soil Science 73(6), e13315, https://doi.org/10.1111/ejss.13315.

Probst, P., Boulesteix, A. L. (2017): To tune or not to tune the number of trees in random forest. Journal of Machine Learning Research 18, 1–18, https://doi.org/10.48550/arXiv.1705.05654.

Probst, P., Wright, M. N., Boulesteix, A. L. (2018): Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9(3), e1301, https://doi.org/10.1002/widm.1301.

Mishra, A. K, Singh, V. P. (2010): A review of drought concepts. Journal of Hydrology 391(1–2), 202–216, https://doi.org/10.1016/j.jhydrol.2010.07.012.

Rahmati, O., Falah, F., Dayal, K. S., Deo, R. C., Mohammadi, F., Biggs, T., Moghaddam, D. D., Naghibi S. A., Bui, D. T. (2019): Machine learning approaches for spatial modeling of agricultural droughts in the south-east region of Queensland Australia. Science of the Total Environment 699, 134230, https://doi.org/10.1016/j.scitotenv.2019.134230.

Rahmati, O., Panahi, M., Kalantari, Z., Soltani, E., Falah, F., Dayal, K. S., Mohammadi, F., Deo, R. C., Tiefenbacher, J., Bui, D. T. (2020): Capability and robustness of novel hybridized models used for drought hazard modeling in southeast Queensland, Australia. Science of the Total Environment 718, 134656, https://doi.org/10.1016/j.scitotenv.2019.134656.

Shashank, S., Pandey, A. C. (2015): Delineation of groundwater potential zone in hard rock terrain of India using remote sensing, geographical information system (GIS) and analytic hierarchy process (AHP) techniques. Geocarto International 30(4), 402–421, https://doi.org/10.1080/10106049.2014.894584.

Simpson, E. H. (1951): The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society: Series B (Methodological) 13(2), 238–241, https://doi.org/10.1111/j.2517-6161.1951.tb00088.x.

Thomas, T., Jaiswal, R. K., Galkate, R., Nayak, P. C., Ghosh, N. C. (2016): Drought indicators-based integrated assessment of drought vulnerability: a case study of Bundelkhand droughts in central India. Natural Hazards 81, 1627–1652, https://doi.org/10.1007/s11069-016-2149-8.

Wilhelmi, O. V., Wilhite, D. A. (2002): Assessing vulnerability to agricultural drought: a Nebraska case study. Natural Hazards 25, 37–58, https://doi.org/10.1023/A:1013388814894.

Yang, M., Mou, Y., Meng, Y., Liu, S., Peng, C., Zhou, X. (2020): Modeling the effects of precipitation and temperature patterns on agricultural drought in China from 1949 to 2015. Science of the Total Environment 711, 135139, https://doi.org/10.1016/j.scitotenv.2019.135139.