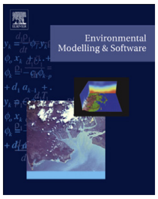
Environmental Modelling & Software 101 (2018) 1–9

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft



Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation

Hanna Meyer [a,*], Christoph Reudenbach [a], Tomislav Hengl [b], Marwan Katurji [c], Thomas Nauss [a]

[a] *Faculty of Geography, Philipps-University Marburg, Deutschhausstr. 10, 35037 Marburg, Germany*
[b] *ISRIC — World Soil Information, P.O. Box 363, 6700 AJ Wageningen, The Netherlands*
[c] *Center for Atmospheric Research, University of Canterbury, Private Bag 4800, Christchurch 8020, New Zealand*

ARTICLE INFO

Article history:
Received 16 October 2017
Accepted 2 December 2017
Available online 14 December 2017

Keywords:
Cross-validation
Feature selection
Over-fitting
Random forest
Spatio-temporal
Target-oriented validation

ABSTRACT

Importance of target-oriented validation strategies for spatio-temporal prediction models is illustrated using two case studies: (1) modelling of air temperature (T_{air}) in Antarctica, and (2) modelling of volumetric water content (VW) for the R.J. Cook Agronomy Farm, USA. Performance of a random *k*-fold cross-validation (CV) was compared to three target-oriented strategies: Leave-Location-Out (LLO), Leave-Time-Out (LTO), and Leave-Location-and-Time-Out (LLTO) CV. Results indicate that considerable differences between random *k*-fold ($R^2 = 0.9$ for T_{air} and 0.92 for VW) and target-oriented CV (LLO $R^2 = 0.24$ for T_{air} and 0.49 for VW) exist, highlighting the need for target-oriented validation to avoid an overoptimistic view on models. Differences between random *k*-fold and target-oriented CV indicate spatial over-fitting caused by misleading variables. To decrease over-fitting, a forward feature selection in conjunction with target-oriented CV is proposed. It decreased over-fitting and simultaneously improved target-oriented performances (LLO CV $R^2 = 0.47$ for T_{air} and 0.55 for VW).

1. Introduction

Machine learning algorithms are well established in environmental sciences (Lary et al., 2016; Kanevski et al., 2009) and find application in a variety of fields as for example mapping of land cover (Ludwig et al., 2016; Gislason et al., 2006), vegetation characteristics (Lehnert et al., 2015; Verrelst et al., 2012) and soil properties (Gasch et al., 2015; Lieβ et al., 2016) as well as in geomorphological (Messenzehl et al., 2017; Micheletti et al., 2014) or climatological (Kühnlein et al., 2014; Hong et al., 2004; Meyer et al., 2016a; Appelhans et al., 2015) studies. Most of the applications focus on static spatial predictions and are not aiming at estimating a certain variable simultaneously in space and time. However, though machine learning algorithms are still rarely applied in spatio-temporal models, the number of applications is increasing (Gokaraju et al., 2011; Gasch et al., 2015; Appelhans et al., 2015; Meyer et al., 2016b; Ho et al., 2014; Jing et al., 2016; Ke et al., 2016; Lary et al., 2014).

Machine learning algorithms in space-time applications learn from spatio-temporal observations to predict a certain variable for unknown locations and for an unknown point in time (within a defined model domain) allowing a monitoring of the environmental variable. The term *"prediction"*, in this context, should not to be confused with *"forecasting"* as most of the models are not aiming at predicting into the future but rather focus on predicting in past or present times as well as in space. In contrast to model-based geostatistics (Diggle and Ribeiro, 2007) as for example (co-) kriging, where one needs sufficiently distributed information on the variable at question for each interpolation time-step, spatio-temporal prediction models link a set of independent variables to the response (i.e. the variable in question) and only use those independent variables for the subsequent spatio-temporal prediction application. A typical example of spatio-temporal prediction models in environmental science might be the estimation of soil properties as done by Gasch et al. (2015). In this example, soil properties (volumetric water content, soil temperature and bulk electrical conductivity) are predicted in space and time on the basis of a machine learning model which is developed from a

[*] Corresponding author.
E-mail address: hanna.meyer@geo.uni-marburg.de (H. Meyer).

https://doi.org/10.1016/j.envsoft.2017.12.001
1364-8152/© 2017 Elsevier Ltd. All rights reserved.

variety of spatial, temporal and spatio-temporal predictor variables as well as *"ground truth"* observations taken from data loggers.

Studies by Gasch et al. (2015) and Meyer et al. (2016a) have shown that the estimated performance of such models highly depends on the validation strategy: in both cases high differences between the performance estimated by a random test subset of the total dataset and the performance estimated by a Leave-Location-Out (LLO) Cross-Validation (CV) have been reported. LLO CV means that models are repeatedly trained by leaving the data from one location or a group of locations (i.e. climate stations, data loggers) out and using the respective held back data for model validation. The differences between a random subset validation (lower error estimates) and LLO CV (higher error estimates) strongly suggest spatial over-fitting as the models can very well predict on subsets of the time series of the locations used for training, but fail in the prediction of unknown locations. The prediction on unknown locations, however, is in most cases the major task of such models. The LLO CV error must therefore be considered as the decisive performance indicator of spatial as well as spatio-temporal models. Similarly, spatio-temporal models have a risk of temporal over-fitting which needs to be assessed by Leave-Time-Out (LTO) CV (Gudmundsson and Seneviratne, 2015). However, it is these *"target-oriented"* validation strategies that focus on the model performance in the context of unknown space or unknown time steps that are not yet fully prevailed in literature. This is especially a problem as case studies ignoring the spatio-temoral dependence in the data have to be considered too optimistic (Roberts et al., 2017). Even though LLO and LTO CV are used in some studies on spatial and spatio-temporal models (Ho et al., 2014; Gudmundsson and Seneviratne, 2015; Ruβ and Brenning, 2010; Meyer et al., 2017b; Brenning et al., 2012; Micheletti et al., 2014), random *k*-fold CV, where the dataset is randomly partitioned into folds, is still considered common practice (Ke et al., 2016; Messenzehl et al., 2017; Lieβ et al., 2016; Ludwig et al., 2016).

How to address spatial or spatio-temporal over-fitting in view to improved model selections? Over-fitting in machine learning models (when applied to spatial data) most likely happens due to poor representation of spatio-temporal sampling in predictor variable spaces. Hence, carefully selecting and interpreting predictor variables is a logical remedy for improving performance of spatial models. Many spatio-temporal prediction studies use auxiliary predictor variables which describe the properties of the location (e.g. elevation, slope, soil type, spatial coordinates). These variables vary in space but not in time which means that each station has a unique combination of static variables. We hypothesize hence that:

1. These temporally static variables are prone to over-fitting. Combinations of unique properties for each location are quasi comparable to a unique ID of the locations which is then used as predictor. Using such variables, the model is able to fit general characteristics of the individual time series.
2. Variables that lead to over-fitting can be automatically identified and removed using a feature selection method that accounts for the target-oriented performance.
3. Excluding misleading variables from the models does not only decrease over-fitting but also leads to improved target-oriented model performances.

Feature selection is an intuitive solution to reduce the number of variables to the most important ones. However, the commonly used method for feature selection, Recursive Feature Elimination (RFE) (see e.g. Brungard et al., 2015; Meyer et al., 2017a, b; Ghosh and Joshi, 2014; Stevens et al., 2013; in the field of environmental mapping), relies on variable importance scores which are calculated using solely the training subset (Kuhn and Johnson, 2013). If a variable leads to considerable over-fitting, it has a high importance in the models. Therefore, this variable will be selected as important variable in the RFE process and is not removed regardless of a resulting high LLO CV error. Alternative approaches for detecting the over-fitting variables are hence required.

We consider two published case studies to demonstrate the effect of different validation strategies, the risk of spatial or spatio-temporal over-fitting as well as the potential of feature selection algorithms to minimize the degree of over-fitting. To estimate the degree of over-fitting, we compare the results of a random *k*-fold CV with the results of the target-oriented validation strategies LLO, LTO and Leave-Location-and-Time-out (LLTO) CV. We then compare the RFE method with a newly proposed forward feature selection (FFS) method that works in conjunction with target-oriented performance to identify and remove variables that lead to over-fitting. As machine learning algorithm, the well-known Random Forest algorithm (Breiman, 2001) was applied as it appeals to a large community of users. We implement all steps of data analysis and modelling in the R environment for statistical programming (R Core Team, 2016). Most of the analysis is based on the caret package (Kuhn, 2016) that implements a wrapper to the Random Forest algorithm being used and provides functionality for data splitting and CV. All newly produced R functions and modelling steps are fully documented in https://github.com/environmentalinformatics-marburg/CAST.

2. Case studies and description of the datasets

2.1. *Case study I: modelling air temperature in Antarctica*

The first case study follows the approach of Meyer et al. (2016a) to spatio-temporally predict T_{air} in Antarctica based on LST data from the Moderate Resolution Imaging Spectroradiometer (MODIS) and auxiliary predictor variables. The dataset as it was used in the present study consists of 30666 hourly air temperature measurements from 32 weather stations distributed over Antarctica for the year 2013. The T_{air} values range from -78.40°C to 5.76°C with an average of -27.64°C and a standard deviation of 17.26°C.

Beside of MODIS based LST as a spatio-temporal predictor variable, several auxiliary spatial predictor variables were used that basically describe the terrain. In addition, a number of predictor variables that remain spatially constant but vary in time were used as temporal predictor variables. See Table 1 for the full list of predictors used in this study and Meyer et al. (2016a) for further information on the dataset.

2.2. *Case study II: modelling volumetric water content of the "Cookfarm", USA*

The second case study bases on the dataset applied in Gasch et al. (2015) to predict soil properties in 3D+time and can be freely accessed from the GSIF package in R. The research site of this case study is the R.J. Cook Agronomy Farm which is a 37 ha sized long-term agroecosystem research site in the Palouse region in the USA and operated by the Washington State University. The final dataset as prepared for this study consists of daily VW measurements from the years 2011–2013 taken by 5TE sensors (Decagon Devices, Inc., Pullman, Washington) initially installed in five depth

Table 1
Predictor variables used within the two case studies with their dimension and resolution (res.). LST — Land Surface Temperature as measured by MODIS, Sensor - either MODIS Terra or Aqua, Ice — Ice covered ground or not, DEM - Digital elevation model, TWI — SAGA wetness index, NDRE.M — Normalized Difference Red Edge Index (mean), NDRE.sd - Normalized Difference Red Edge Index (s.d.), Bt — Occurrence of Bt horizon, BLD — Bulk density of soil, PHI — Soil pH, Precip_cum — Cumulative precipitation in mm, MaxT_wrcc - Maximum measured temperature, MinT_wrcc — Minimum measured temperature, Crop — Crop type. See also Meyer et al. (2016a) and Gasch et al. (2015) for further description.

Case Study	Predictor	Dimension	Spatial res.	Temporal res.
T_{air} Antarctica	LST	2D+t	1km	instantaneous
	DEM	2D	1 km	–
	Aspect	2D	1 km	–
	Slope	2D	1 km	–
	Skyview	2D	1 km	–
	Ice	2D	1 km	–
	Sensor	(2D+t)	(1 km)	(instantaneous)
	Season	t	–	3 months
	Time	t	–	hour
	Month	t	–	1 month
VW Cookfarm	DEM	2D	10 m	–
	TWI	2D	10 m	–
	NDRE.M	2D	10 m	–
	NDRE.Sd	2D	10 m	–
	Bt	2D	10 m	–
	BLD	2D	10 m	–
	PHI	2D	10 m	–
	Precip_cum	t	–	1 day
	MaxT_wrcc	t	–	1 day
	MinT_wrcc	t	–	1 day
	Cdayt	t	–	1 day
	Crop	2D+t	10 m	1 year

(0.3, 0.6, 0.9, 1.2, and 1.5 m) at 42 locations within the study site. In this study we only focus on two dimensions plus time and limited the dataset to the depth of 0.3 m. The dataset then contained 33397 training samples. VW ranged from 0.093 m^3/m^3 to 0.613 m^3/m^3 with an average of 0.265 m^3/m^3 and a standard deviation of 0.076 m^3/m^3.

The covariables available from the research dataset that were used in this study as potential predictors to predict VW are a number of spatially continuous variables describing the terrain. Further, temporal variables as for example climate properties measured from the nearest meteorological station were used. See Table 1 for the full list of predictors used in this study and Gasch et al. (2015) for further information on the dataset.

3. Methods

3.1. Random forest algorithm

Random Forest bases on the concept of regression and classification trees, i.e. a series of nested decision rules for the predictors that determine the response. It repeatedly builds trees from random samples of the training data with each tree is a separate model of the ensemble. The estimations of all trees are finally averaged to produce the final estimate (Breiman, 2001). To overcome correlation between trees, only a subset of predictors (mtry) is randomly selected at each split. The best predictor from the random subset is used at the respective split to partition the data. mtry is considered as a hyperparameter that needs to be tuned for a respective dataset in order to obtain an optimal trade-off between under- and over-fitting of the data. For a further description of Random Forest, see Breiman (2001); James et al. (2013) and Kuhn and Johnson (2013).

In this study, the Random Forest implementation of the randomForest package (Liaw and Wiener, 2002) in R was applied and accessed via the caret package (Kuhn, 2016). Throughout the study, each Random Forest model consisted of 500 trees after no increase of performance could be observed using a higher number of trees. mtry was tuned for each value between two and the respective number of predictor variables.

3.2. Validation strategies

To test the model performance on random subsets of the total datasets, a commonly used random 10-fold CV was used. Therefore, the data was split into 10 equally sized folds. Data splitting was done by stratified random sampling that ensures that the distribution of the response variable in each fold equals the distribution of the entire dataset. Models were then repeatedly trained by using the data of all except one fold and testing the model performance using the held-back data (Fig. 2). In order to quantify the performance of the models using *"target-oriented"* validation strategies, the performance in view to the following criteria was tested (Fig. 1).

1. predict on unknown locations, tested by Leave-Location-Out Cross-Validation (LLO CV)
2. predict on unknown points in time, tested by Leave-Time-Out Cross-Validation (LTO CV)
3. predict on unknown locations and unknown points in time, tested by Leave-Time-and-Location-Out Cross-Validation (LLTO CV)

Therefore, the dataset was split into folds again, but this time each fold left the data of complete locations (LLO) or time steps (LTO) or locations as well as time steps (LLTO) out (Fig. 2). For both case studies, the location of the data loggers defined a location and the dataset was split into 10 folds with respect to these locations. For LTO, the day of the year was used as splitting criterion for T_{air} Antarctica. For VW Cookfarm, data from more than one year was available allowing that individual months of each year could be left out for validation (12 months × 3 years = 36 unique time steps). Again, the data was split into 10 folds by leaving complete time steps out.

For all target-oriented validation strategies, the procedure was comparable to the random *k*-fold validation (which gives a biased estimate of prediction performance): models were repeatedly trained by using the data of all except one fold and testing the model performance for the held-back data. Over-fitting of the model in space and time was then quantified by comparing the random 10-fold CV results with the target-oriented validation results.

3.3. Feature selection

With the aim to remove predictors that are counterproductive in view to the target-oriented performance, we tested a RFE algorithm as well as a FFS algorithm that works in conjunction with target-oriented validation (Fig. 1). We used LLO and LLTO CV as target-oriented validation strategies as the ability of the model to predict on unknown locations was of upmost importance for both case studies.

RFE relies on variable importance scores that are calculated during the initial random forest model training. The algorithm successively removes the least important variables to find the best performing set (see Kuhn and Johnson, 2013; for further details). In this study, we used the RFE implementation from the caret package

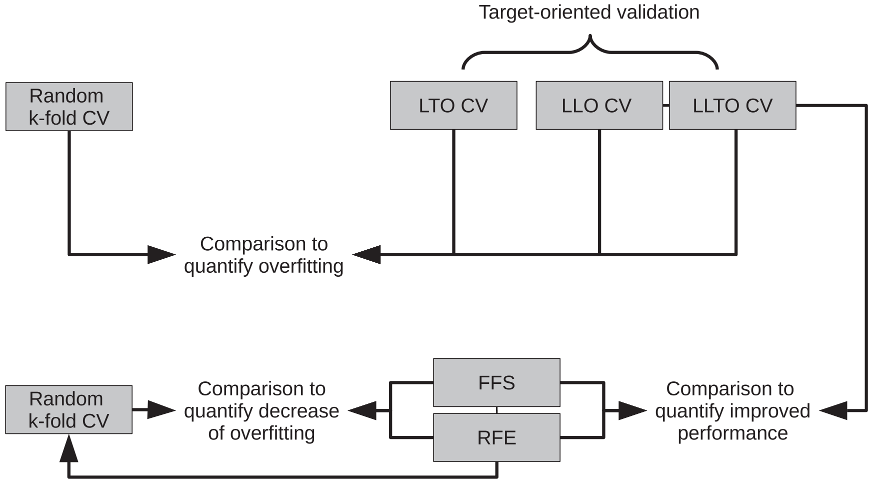

Target-oriented validation
Random k-fold CV
LTO CV
LLO CV
LLTO CV
Comparison to quantify overfitting
Random k-fold CV
Comparison to quantify decrease of overfitting
FFS
RFE
Comparison to quantify improved performance

Fig. 1. Schematic overview of validation strategies considered in this study. Leave-Location-Out (LLO), Leave-Time-Out (LTO) and Leave-Location-and-Time-Out (LLTO) cross-validations (CV) were used as target-oriented strategies. LLO and LLTO CV were used in conjunction with recursive feature elimination (RFE) and forward feature selection (FFS) to reduce spatial over-fitting and its impact.

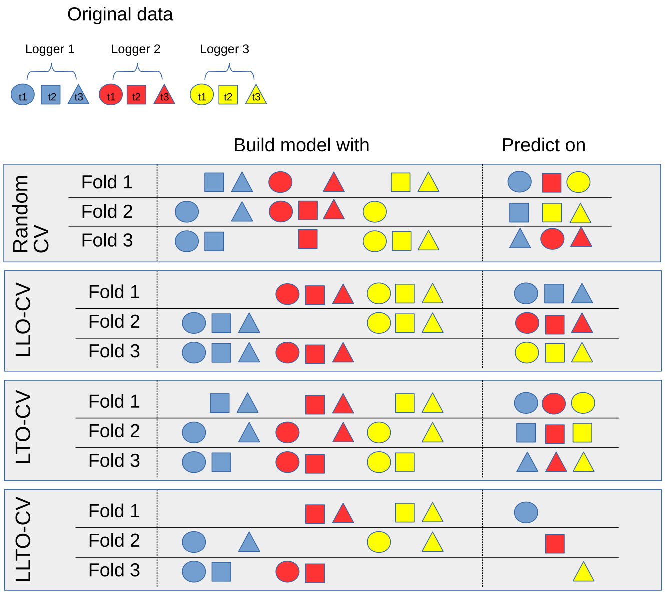


Fig. 2. Schematic overview of the validation strategies 3-fold random cross-validations (CV), Leave-Location-Out (LLO), Leave-Time-Out (LTO) and Leave-Location-and-Time-Out (LLTO) CV. At this example, the original data base on three data loggers as indicated by color, each measured at three points in time (t) as indicated by shape. CV bases on three folds. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

(Kuhn, 2016). As outlined above, we assume that RFE is not a helpful approach to overcome spatio-temporal over-fitting as variables are not ranked according to target-oriented performance. As an alternative approach, we developed and implemented a FFS algorithm in R (Algorithm 1). The algorithm first trains models (i.e. Random Forest) of all possible 2-variable combinations of the total set of predictor variables. The best initial model in view to target-oriented performance is kept. The number of predictor variables is then iteratively increased. The improvement of the model is tested for each additional predictor using target-oriented CV. The process stops when none of the remaining variables decreases the error of the currently best model. The algorithm therefore fits a maximum of $2*(n-1)^2/2$ models (e.g. 81 models when 10 predictors are considered).

```
for each resampling iteration do
    Partition the data into training and test data
    Tune and train models using all possible 2-variable combinations
    Predict on test data and calculate model performance
end
Keep the best performing 2-variable model ($model_{best}$)
for each additional number of variables i, i=3...N do
    for each remaining variable $V_R$ do
        for each resampling iteration do
            Partition the data into training and test data
            Tune and train models using the variables of $model_{best}$ and $V_R$
            Predict on test data and calculate model performance
        end
    end
    if $mean(error\ of\ model_i) > mean(error\ of\ model_{best})$ then
        break
    end
    Keep the best performing i-variable model ($model_{best}$)
end
```

Algorithm 1: Pseudo-code description (similar to the ones from the caret package) for the FFS algorithm. Resampling in this study bases either on LLO or LLTO.

4. Results and discussions

4.1. Target-oriented validation

For both case studies, a random *k*-fold CV showed a high performance with only low differences between observed and predicted values, indicating a nearly "perfect fit" of the data (model T_{air}/VW01 in Table 2). However, in view to unknown locations (LLO CV), the performance decreased considerably (models T_{air}/VW02 compared to models T_{air}/VW01 in Table 2). This means that the

Table 2
Regression statistics between observed and predicted values of air temperature (T_{air}) and volumetric water content (VW) based on cross-validation (CV). Models were validated using random k-fold or using target-oriented Leave-Location-Out (LLO), Leave-Time-Out (LTO) and Leave-Location-and-Time-Out (LLTO) CV. Recursive feature elimination (RFE) and the newly proposed forward feature selection (FFS) were tested. Performance measures are mean error (ME), mean absolute error (MAE), root-mean-square-error (RMSE) and coefficient of determination (R^2). Bold numbers indicate the decisive objective error estimates after misleading variables were removed by FFS. Compare target-oriented CV without feature selection to random k-fold CV to estimate over-fitting. Compare LLO and LLTO CV using RFE or FFS to estimate the increase of performance compared to LLO and LLTO CV without feature selection. Note that the random CV performance is only provided for comparison but cannot be regarded as a meaningful measure.

Model	CV	Feature Select.	ME	MAE	RMSE	R^2
T_{air}01	random	none	0.016	4.155	5.556	0.899
T_{air}02	LLO	none	0.068	12.178	15.850	0.244
T_{air}03	LTO	none	0.017	4.244	5.665	0.894
T_{air}04	LLTO	none	0.236	12.164	15.807	0.246
T_{air}05	LLO	RFE	0.011	10.353	13.647	0.400
T_{air}06	random	variables of T_{air}05	0.025	9.113	12.021	0.519
T_{air}07	LLO	FFS	**0.072**	**9.756**	**12.564**	**0.474**
T_{air}08	random	variables of T_{air}07	0.000	8.602	11.157	0.583
T_{air}09	LLTO	RFE	0.405	10.251	13.416	0.413
T_{air}10	random	variables of T_{air}09	0.025	9.113	12.021	0.519
T_{air}11	LLTO	FFS	**0.253**	**9.658**	**12.387**	**0.485**
T_{air}12	random	variables of T_{air}11	−0.001	8.601	11.156	0.583
*VW*01	random	none	0.00	0.016	0.023	0.919
*VW*02	LLO	none	−0.002	0.041	0.054	0.488
*VW*03	LTO	none	−0.001	0.024	0.035	0.794
*VW*04	LLTO	none	−0.007	0.040	0.050	0.500
*VW*05	LLO	RFE	−0.002	0.041	0.055	0.475
*VW*06	random	variables of *VW*05	0.000	0.014	0.021	0.931
*VW*07	LLO	FFS	**0.000**	**0.037**	**0.051**	**0.552**
*VW*08	random	variables of *VW*07	0.000	0.036	0.049	0.580
*VW*09	LLTO	RFE	−0.007	0.040	0.050	0.502
*VW*10	random	variables of *VW*09	0.00	0.015	0.022	0.926
*VW*11	LLTO	FFS	**−0.004**	**0.039**	**0.051**	**0.499**
*VW*12	random	variables of *VW*11	0.00	0.036	0.049	0.580

model was generally less able to predict beyond the location of the training data compared to what might have been expected regarding the random k-fold CV error. The ability of the T_{air} model to predict the outcome for an unknown day within the temporal model domain of 2013 remained high (model T_{air}03 in Table 2). Thus, in view to unknown locations and unknown days (model T_{air}04 in Table 2), the error was comparable to the LLO CV error. Uncertainties in view to unknown locations were the major source of error. The temporal error had more effect on the VW Cookfarm example where complete months were left out for validation (model VW03 and VW04 in Table 2).

Since the differences between random k-fold CV and target-oriented CV are noticeably high, the results highlight the need to perform CV in view to the model target in order to draw meaningful conclusions. If the aim is to map the response variable, one must consider LLO CV as decisive error indicator as the random CV error can lead to considerable misinterpretations of the model performance. Especially when the model is to be applied on unknown years, the potential of the model to predict beyond the years used for model training must also be considered. In this case, LLTO CV can assess the error in both, space and time, however, the number of validation data decreases as the overlap between LLO and LTO is used. This causes the results to be less robust compared to a separate view on LLO and LTO CV were more data are available for testing.

4.2. Detecting over-fitting

As the models performed well on random subsets of the entire datasets (random k-fold CV) but had high errors when faced with unknown locations, spatial over-fitting must be suspected for both case studies. The model could only lead to high performances when information about a respective location went into model training. Therefore, the model was over-fitting in space as only locations used for training could reliably be predicted by the model. Subsequently, also LLTO CV showed high errors, though temporal over-fitting only slightly contributed to that error in case of the T_{air} Antarctica example. In this case study, over-fitting in time was a minor issue, at least on the considered time scale (days). In the case study of VW Cookfarm, the time scale used for data splitting was months of the individual years. Considering these larger time scales that were left out, the model performance decreased compared to the random k-fold CV performance ($R^2 = 0.79$ compared to 0.92, see VW03,01 in Table 2). Thus, temporal over-fitting must be assumed in addition to spatial over-fitting as only months that went into model training could reliably be predicted by the model.

4.3. Reducing over-fitting and improving model performances

To decrease the impact of over-fitting, RFE and the newly designed FFS were compared. On the first sight, RFE reduced over-fitting in the T_{air} Antarctica example, getting obvious in lower differences between random k-fold CV and target-oriented CV (Fig. 3a, model T_{air}05 compared to T_{air}06 as well as T_{air}09 compared to T_{air}10 in Table 2). This pattern, however, could not be supported by the VW Cookfarm example, where the differences between random k-fold CV and target-oriented CV remained equally high (Fig. 3b, model VW05 compared to VW06 as well as VW09 compared to VW10 in Table 2). In fact, this was the expected pattern as the variable importance ranking within the RFE is based on internal importance estimates (Fig. 4) without consideration of the importance in view to target-oriented errors.

The explanation for the effect shown in the T_{air} example lies in the ranking of the variables (Fig. 4a): Among the most important variables were apparently those that do not lead to an over-fitting. Only the top three variables were selected by the RFE (”season”, ”time” and ”LST″, see Table 1 for explanation) that could in this example lead to a reduced effect of over-fitting. Including just one additional variable (in this case ”aspect” as this was the variable rated as next important, see Fig. 4) was recognised as counterproductive by the RFE. However, the example of VW Cookfarm demonstrates that this pattern is rather chance than a systematic ability of the RFE design to remove over-fitting variables. In the VW Cookfarm example the variables that were ranked as important led to over-fitting so that the RFE could not decrease this problem by removing least important variables. Over-fitting in this example is generated because the variables were not ranked according to their target-oriented importance within the models. In fact, the RFE algorithm kept all except three variables thus it yielded the best performance using nearly the full set of predictors which, however, could not remove over-fitting.

The FFS algorithm, in contrast, could reliably reduce the differences between random k-fold CV and LLO as well as LLTO CV in both case studies: when the respective less-variable model was validated with random k-fold CV, the differences to the LLO as well as LLTO CV error decreased (Fig. 3, model VW/T_{air}07,11 compared to VW/T_{air}10,12 in Table 2). This shows that removing misleading variables decreased the problem of spatial over-fitting. In the case study of T_{air} Antarctica, it suggested the combination of ”season”, ”ice”, ”LST″, ”sensor”, ”month” as necessary variables and rated all others as counterproductive. For VW Cookfarm, the variables ”Precip_cum”, ”cdayt”, ”MaxT_wrcc”, ”MinT_wrcc”, ”Crop” were suggested to yield optimal results in view to LLO as well as LLTO CV.

The variables that were rated as counterproductive and have

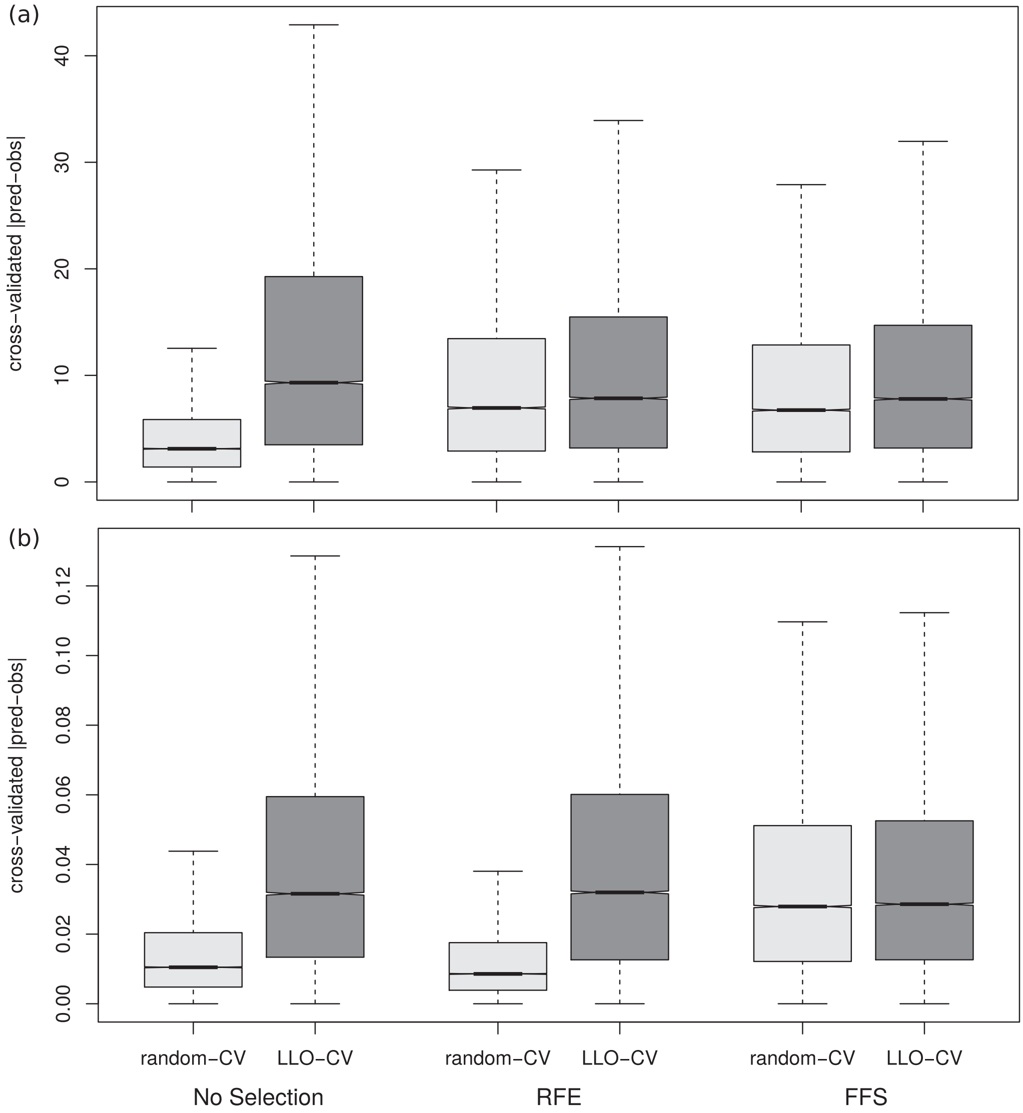


Fig. 3. Differences in the Leave-Location-Out (LLO) cross-validation (CV) performance of a) the air temperature (T_{air}) estimations and b) the volumetric water content (VW) estimations using different feature selection strategies. The effect of a Recursive feature elimination (RFE) and the newly proposed forward feature selection (FFS) are compared. The variables selected by RFE for the case study of T_{air} Antarctica were "season", "time", "LST". FFS selected "month", "ice", "LST", "season", "sensor". For the case study of VW Cookfarm all potential predictors except "Bt", "TWI", "MinT_wrcc" were selected by the RFE. FFS selected "MaxT_wrcc", "cdayt", "Precip_cum", "Crop", "MinT_wrcc". See Table 1 for further explanations on the variables.

been removed during FFS were spatially continuous but temporally constant variables. The only exception was the variable "ice" for T_{air} Antarctica, which however, features a high overlap between logger locations, thus it does not allow for a unique "pointer" on the individual logger locations. Especially in the case study of T_{air} Antarctica, the temporally static variables formed a distinct "pointer" on the individual logger locations, as each logger location featured unique combinations of the spatial variables (i.e. unique combinations of slope, aspect, altitude). Therefore, these variables are, in combination, comparable to an "ID" for the loggers that was then used as predictor. ID-like predictors enable the algorithms to access individual characteristics of the time series of the loggers which in turn leads to a misinterpretation of such variables: these variables are associated with logger-specific patterns that cover the true underlaying relations between these predictors and the response. This suspicion is supported by a high internal importance of such variables within the models (Fig. 4) especially in the VW Cookfarm example (e.g. NDRE.M+BLD+PHI) but a removal of these variables during the FFS. Under these considerations, the behaviour of the RFE to reduce the impact of over-fitting in the T_{air} Antarctica example becomes understandable: as the top ranked variables "season", "time" and "LST" are not prone to spatial over-fitting, the RFE could yield best results using only these three variables. If an over-fitting variable was amongst the top two variables, over-fitting could not have been resolved and the differences between random k-fold and target-oriented CV errors stayed high as in the VW Cookfarm example.

Removing counterproductive variables using FFS did not only lead to reduced over-fitting but also to improved target-oriented performances (Figs. 3 and 5, Table 2). This is especially obvious for the T_{air} Antarctica data where the LLO CV R^2 increased from 0.24 to 0.47 (model T_{air}07 compared to T_{air}02 in Table 2, Figs. 3a, and 5a). The patterns for LLTO CV were the same (model T_{air}11 compared to T_{air}04 in Table 2, Fig. 5a). Also in the VW Cookfarm example FFS led to an increased LLO performance, though the effect was less strong compared to the T_{air} Antarctica data (Fig. 5b). The LLO CV R^2

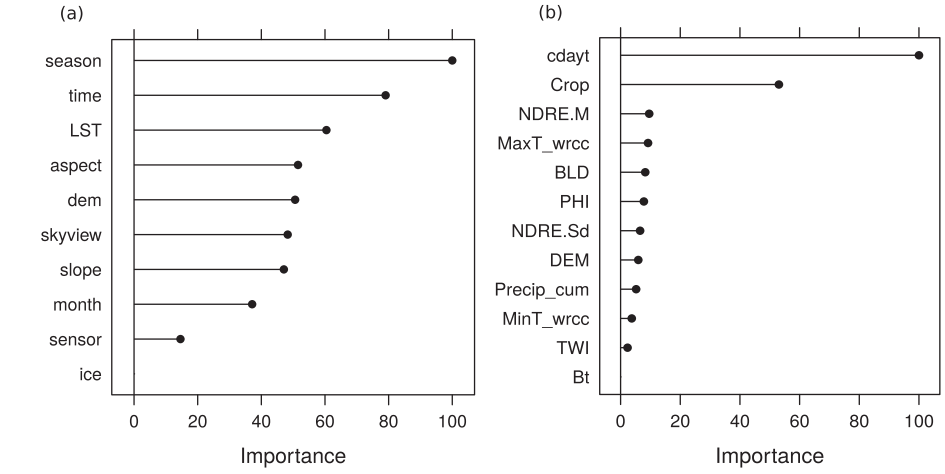


Fig. 4. Relative scaled importance of the predictor variables within the Random Forest models for the case study of (a) T_{air} Antarctica and (b) VW Cookfarm. See Table 1 for further explanations on the variables.

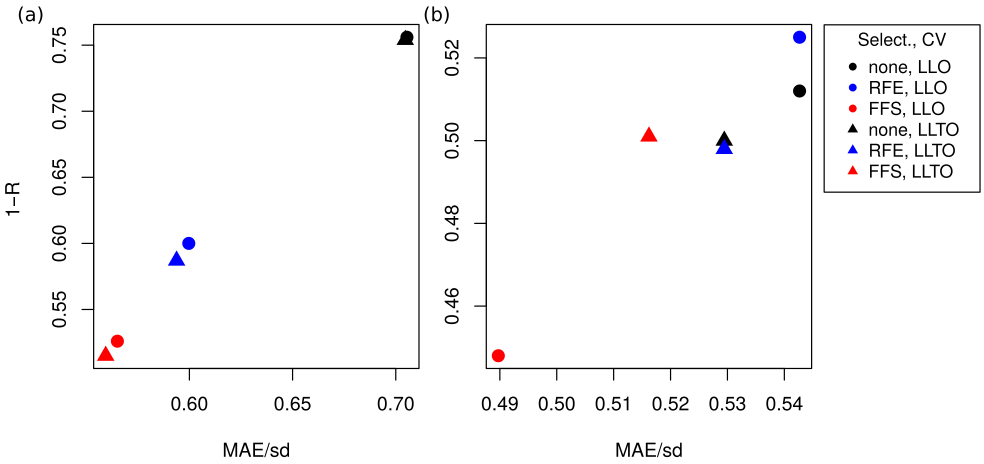


Fig. 5. Differences in the Leave-Location-Out (LLO) and Leave-Location-and-Time-Out (LLTO) cross-validation (CV) performance using no feature selection, a recursive feature elimination (RFE) and the newly proposed forward feature selection (FFS) of the a) air temperature (T_{air}) Antarctica models and b) the volumetric water content (VW) Cookfarm models. Performance is indicated using the mean absolute error (MAE) divided by the standard deviation (sd) of the mean, and the proportion of variation unexplained (1 - R^2). Colors indicate the different feature selection strategies. The shape indicates the CV method being used. Performance increases from the upper right corner towards the lower left corner. It is shown that models using no feature selection generally have the lowest performance and models using FFS have the highest performance. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

increased from 0.49 to 0.55 (model VW02 compared to VW07 in Table 2) using only the selected variables. For the LLTO CV error, the FFS did not resulted in an improved model performance (model VW04 compared to VW11 in Table 2, Fig. 5b) though over-fitting could be significantly removed (Fig. 3b). Obviously removing misinterpreted variables could not improve the performance which suggests that the potential of the variables to predict beyond the training locations and months is depleted. However, this model is now more robust as only a small subset of the initial variables are used and over-fitting could be reduced.

Though FFS is time consuming, it is able to automatically detect and remove variables that are counterproductive in view to the target. The computation time can be decreased by thorough preselection of potential predictors in view to their effect in space and time to avoid ID-like pointers on individual locations or time steps. Considering the potential of FFS as shown in this study to remove counterproductive variables in view to a target-oriented performance, it is likely that the algorithm is able to improve a variety of published models beside of the two case studies (Meyer et al., 2016a; Gasch et al., 2015). As an example, Langella et al. (2010); Shi et al. (2015) and Janatian et al. (2017) used latitude and longitude as predictors which are prone to create an ID of the locations used for training.

The focus of this study was on spatio-temporal models, however, most of the findings apply for purely spatial models as well. This is supported by the studies of e.g. Micheletti et al. (2014) and Roberts et al. (2017) who left spatial units out for validation and yielded less optimistic results compared to a random k-fold CV, thus spatial over-fitting is indicated. Also Li et al. (2011) included latitude and longitude as predictors in a purely spatial model and observed linear features in the resulting map. If such models are validated with random k-fold CV, a statistically good fit is feigned but spatial over-fitting occurs as a consequence of the misinterpretation of certain variables. In such applications, the proposed FFS in conjunction with target-oriented validation (in this case leave-spatial-unit-out CV) can improve the model results and will produce more robust results.

5. Conclusions

This study addressed the widely ignored problem of dependencies caused by the nature of spatio-temporal data. It aimed at demonstrating the effect of target-oriented validation and at finding a solution to detect and reduce spatial over-fitting. For this we used two previously published case studies. We discovered high differences between random *k*-fold and target-oriented CV: the random *k*-fold CV R^2 of the T_{air} Antarctica study was 0.90, contrasting to a LLO CV R^2 of 0.24 and the random *k*-fold CV R^2 of the VW Cookfarm study was 0.92 compared to a LLO CV R^2 of 0.49. This shows that errors estimated with a standard random *k*-fold CV can considerably deviate from target-oriented error estimates which highlights the clear need for target-oriented validation to avoid an overoptimistic view on results.

We further hypothesized that the observed patterns of spatio-temporal over-fitting are caused by temporally constant predictors (e.g. elevation, slope, …) that act in conjunction with each other like an ID. This occurs when locations used for model training have unique spatial properties. It appears that the models of both case studies were able to learn general characteristics of the time series of the individual locations. The models were then very well able to predict subsets of the time series (low random *k*-fold CV error), but then failed to predict beyond the training locations (high LL(T)O CV error). To automatically detect and remove variables that lead to over-fitting, we proposed using the FFS algorithm in conjunction with target-oriented validation. By removing the misleading predictors, the FFS was able to automatically reduce spatio-temporal over-fitting which was reflected in similar errors for random *k*-fold CV and target-oriented CV. After refitting the models using the FFS procedure, the LLO CV R^2 was 0.47 for T_{air} Antarctica and 0.55 for VW Cookfarm, hence the proposed method could improve the target-oriented model performance.

In this study, we applied the frequently used Random Forest algorithm. Though other algorithms have not explicitely been tested in this study, the importance of target-oriented validation as well as the risk of spatial over-fitting is independent of the algorithm, and applies to other flexible algorithms as e.g. neural networks, support vector machines or cubist (as indicated by Meyer et al., 2016a) in the same way.

Though predicting environmental variables in space and time remains challenging, validation strategies suggested in this article allow assessing model errors objectively and allow identifying over-fitting. The shown importance of target-oriented validation which is widely underestimated, as well as the risk for considerable spatial overfitting indicate that the newly proposed modelling framework should become common research practice for spatio-temporal prediction modelling. Despite the general opinion that Random Forests (and other flexible algorithms) are insensitive to over-fitting, unfavorable combinations of predictors and/or distribution of the training data in space and time can lead to serious over-fitting effects. In this study, variables that has caused that over-fitting were removed from the models and the model performance has immediately improved. However, certain variables might be misleading but still contain valuable information. How to minimize the over-fitting effect of such variables but still use them in the spatial prediction, remains to be solved. With an increasing application of machine learning for spatio-temporal predictions, further studies and procedures for preventing over-fitting in machine learning applications will hence be increasingly important.

Acknowledgements

This work was partly funded by the Federal Ministry of Education and Research (BMBF) within the IDESSA project (grant no. 01LL1301) which is part of the SPACES-program (Science Partnership for the Assessment of Complex Earth System processes). It was further supported by the Ross Sea Region Terrestrial Data Analysis research program, funded by the Ministry of Business and Innovation, New Zealand (grant no. CO9X1413).

References

Appelhans, T., Mwangomo, E., Hardy, D.R., Hemp, A., Nauss, T., 2015. Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. Spat. Stat. 14 (Part A), 91–113.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Brenning, A., Long, S., Fieguth, P., 2012. Detecting rock glacier flow structures using Gabor filters and IKONOS imagery. Remote Sens. Environ. 125, 227–237.

Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Jr, T.C.E., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239–240, 68–83.

Diggle, P., Ribeiro, P.J., 2007. Model-based Geostatistics. Springer Series in Statistics. Springer.

Gasch, C.K., Hengl, T., Gräler, B., Meyer, H., Magney, T.S., Brown, D.J., 2015. Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3D + T: the Cook Agronomy Farm data set. Spat. Stat. 14 (Part A), 70–90.

Ghosh, A., Joshi, P., 2014. A comparison of selected classification algorithms for mapping bamboo patches in lower Gangetic plains using very high resolution WorldView 2 imagery. Int. J. Appl. Earth Observation Geoinformation 26, 298–311.

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random Forests for land cover classification. Pattern Recognit. Lett. 27, 294–300.

Gokaraju, B., Durbha, S.S., King, R.L., Younan, N.H., 2011. A machine learning based spatio-temporal data mining approach for detection of harmful algal blooms in the gulf of Mexico. IEEE J. Sel. Top. Appl. Earth Observations Remote Sens. 4, 710–720.

Gudmundsson, L., Seneviratne, S.I., 2015. Towards observation-based gridded runoff estimates for Europe. Hydrology Earth Syst. Sci. 19, 2859–2879.

Ho, H.C., Knudby, A., Sirovyak, P., Xu, Y., Hodul, M., Henderson, S.B., 2014. Mapping maximum urban air temperature on hot summer days. Remote Sens. Environ. 154, 38–45.

Hong, Y., Hsu, K.-L., Sorooshian, S., Gao, X., 2004. Precipitation estimation from remotely sensed imagery using an artificial neural network cloud classification System. J. Appl. Meteorology 43, 1834–1853.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning: with Applications in R, first ed. Springer, New York.

Janatian, N., Sadeghi, M., Sanaeinejad, S.H., Bakhshian, E., Farid, A., Hasheminia, S.M., Ghazanfari, S., 2017. A statistical framework for estimating air temperature using MODIS land surface temperature data. Int. J. Climatol. 37, 1181–1194.

Jing, W., Yang, Y., Yue, X., Zhao, X., 2016. A comparison of different regression algorithms for downscaling monthly satellite-based precipitation over north China. Remote Sens. 8, 835.

Kanevski, M., Pozdnukhov, A., Timonin, V., 2009. Machine Learning for Spatial Environmental Data: Theory, Applications and Software, first ed. EPFL Press.

Ke, Y., Im, J., Park, S., Gong, H., 2016. Downscaling of MODIS one kilometer evapotranspiration using Landsat-8 data and machine learning approaches. Remote Sens. 8, 215.

Kuhn, M., 2016. Caret: Classification and Regression Training r package version 6.0-68. https://CRAN.R-project.org/package=caret.

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling, first ed. Springer, New York.

Kühnlein, M., Appelhans, T., Thies, B., Nauss, T., 2014. Precipitation estimates from MSG SEVIRI daytime, nighttime, and twilight data with random forests. J. Appl. Meteor. Climatol. 53, 2457–2480.

Langella, G., Basile, A., Bonfante, A., Terribile, F., 2010. High-resolution space-time rainfall analysis using integrated ANN inference systems. J. Hydrology 387, 328–342.

Lary, D., Faruque, F., Malakar, N., Moore, A., Roscoe, B., Adams, Z., Eggelston, Y., 2014. Estimating the global abundance of ground level presence of particulate matter (PM2.5). Geospatial Health 8, 611–630.

Lary, D.J., Alavi, A.H., Gandomi, A.H., Walker, A.L., 2016. Machine learning in geosciences and remote sensing. Geosci. Front. 7, 3–10.

Lehnert, L.W., Meyer, H., Wang, Y., Miehe, G., Thies, B., Reudenbach, C., Bendix, J., 2015. Retrieval of grassland plant coverage on the Tibetan Plateau based on a multi-scale, multi-sensor and multi-method approach. Remote Sens. Environ. 164, 197–207.

Li, J., Heap, A.D., Potter, A., Daniell, J.J., 2011. Application of machine learning methods to spatial interpolation of environmental variables. Environ. Model. Softw. 26, 1647–1659.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R. News 2, 18–22.

Ließ, M., Schmidt, J., Glaser, B., 2016. Improving the spatial prediction of soil organic carbon stocks in a Complex tropical mountain landscape by methodological specifications in machine learning approaches. PLOS ONE 11, 1–22.

Ludwig, A., Meyer, H., Nauss, T., 2016. Automatic classification of Google Earth images for a larger scale monitoring of bush encroachment in South Africa. Int. J. Appl. Earth Observation Geoinformation 50, 89–94.

Messenzehl, K., Meyer, H., Otto, J.-C., Hoffmann, T., Dikau, R., 2017. Regional-scale controls on the spatial activity of rockfalls (Turtmann Valley, Swiss Alps) – a multivariate modeling approach. Geomorphology 287, 29–45.

Meyer, H., Katurji, M., Appelhans, T., Müller, M.U., Nauss, T., Roudier, P., Zawar-Reza, P., 2016a. Mapping daily air temperature for Antarctica based on MODIS LST. Remote Sens. 8, 732.

Meyer, H., Kühnlein, M., Appelhans, T., Nauss, T., 2016b. Comparison of four machine learning algorithms for their applicability in satellite-based optical rainfall retrievals. Atmos. Res. 169 (Part B), 424–433.

Meyer, H., Kühnlein, M., Reudenbach, C., Nauss, T., 2017a. Revealing the potential of spectral and textural predictor variables in a neural network-based rainfall retrieval technique. Remote Sens. Lett. 8, 647–656.

Meyer, H., Lehnert, L.W., Wang, Y., Reudenbach, C., Nauss, T., Bendix, J., 2017b. From local spectral measurements to maps of vegetation cover and biomass on the Qinghai-Tibet-Plateau: do we need hyperspectral information? Int. J. Appl. Earth Observation Geoinformation 55, 21–31.

Micheletti, N., Foresti, L., Robert, S., Leuenberger, M., Pedrazzini, A., Jaboyedoff, M., Kanevski, M., 2014. Machine learning feature selection methods for landslide susceptibility mapping. Math. Geosci. 46, 33–57.

R Core Team, 2016. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929.

Ruß, G., Brenning, A., 2010. Data mining in precision agriculture: management of spatial information. In: Hüllermeier, E., Kruse, R., Hoffmann, F. (Eds.), Computational Intelligence for Knowledge-based Systems Design: 13th International Conference on Information Processing and Management of Uncertainty, IPMU 2010, Dortmund, Germany, June 28-July 2, 2010. Proceedings. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 350–359.

Shi, Y., Song, L., Xia, Z., Lin, Y., Myneni, R.B., Choi, S., Wang, L., Ni, X., Lao, C., Yang, F., 2015. Mapping annual precipitation across mainland China in the period 2001-2010 from TRMM3B43 product using spatial downscaling approach. Remote Sens. 7, 5849–5878.

Stevens, A., Nocita, M., Tóth, G., Montanarella, L., van Wesemael, B., 2013. Prediction of soil organic carbon at the european scale by visible and near InfraRed reflectance spectroscopy. PLOS ONE 8, 1–13.

Verrelst, J., Muñoz, J., Alonso, L., Delegido, J., Rivera, J.P., Camps-Valls, G., Moreno, J., 2012. Machine learning regression algorithms for biophysical parameter retrieval: opportunities for Sentinel-2 and -3. Remote Sens. Environ. 118, 127–139.