Geoderma 316 (2018) 100–114

Contents lists available at ScienceDirect

Geoderma

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

Spatial prediction of soil water retention in a Páramo landscape: Methodological insight into machine learning using random forest

Carlos M. Guio Blanco[a], Victor M. Brito Gomez[b,c], Patricio Crespo[c], Mareike Ließ[a,*]

[a] Department of Soil System Science, Helmholtz Centre for Environmental Research – UFZ, Halle (Saale), Germany
[b] Department of Geosciences/Soil Physics Division, University of Bayreuth, Bayreuth, Germany
[c] Departamento de Recursos Hídricos y Ciencias Ambientales, Facultad de Ciencias Agropecuarias, Universidad de Cuenca, Cuenca, Ecuador

ARTICLE INFO

Editor: A.B. McBratney

Keywords:
Water retention
Páramo
Random Forest
Validation
Parameter tuning

ABSTRACT

Soils of Páramo ecosystems regulate the water supply to many Andean populations. In spite of being a necessary input to distributed hydrological models, regionalized soil water retention data from these areas are currently not available. The investigated catchment of the Quinuas River has a size of about 90 km^2 and comprises parts of the Cajas National Park in southern Ecuador. It is dominated by soils with high organic carbon contents, which display characteristics of volcanic influence. Besides providing spatial predictions of soil water retention at the catchment scale, the study presents a detailed methodological insight to model setup and validation of the underlying machine learning approach with random forest. The developed models performed well predicting volumetric water contents between 0.55 and 0.9 cm^3 cm^{-3}. Among the predictors derived from a digital elevation model and a Landsat image, altitude and several vegetation indices provided the most information content. The regionalized maps show particularly low water retention values in the lower Quinuas valley, which go along with high prediction uncertainties. Due to the small size of the dataset, mineral soils could not be separated from organic soils, leading to a high prediction uncertainty in the lower part of the valley, where the soils are influenced by anthropogenic land use.

1. Introduction

Páramo ecosystems are found between 11°N to 8°S in the South American Andes, forming a discontinuous belt between Venezuela and northern Peru (Arroyo et al., 2013; Buytaert et al., 2006a). They are providing water-related ecosystem services to Andean communities (Asbjornsen et al., 2017; Viviroli et al., 2007) and are referred to as sentinels for climate change (Dangles et al., 2017). Páramo soils are identified as one of the most important biophysical components in order to maintain hydrological services and understand the ecohydrological functioning of the system (Mosquera et al., 2015; Schneider et al., 2016).

The soils which are described as volcanic ash soils with high organic matter contents (Buytaert et al., 2007) have commonly high water retention values (Buytaert et al., 2006a). Information regarding the spatial heterogeneity of these soils' water retention curves (WRC) in the form of high-resolution maps, including uncertainty estimates, is necessary for understanding, modeling, and management of these ecosystems (e.g. Horta et al., 2014).

The WRC describes the volumetric soil water content (θ) at equilibrium at different matric potentials (from here on, reported as the logarithm of the height of the water column, pF). It is related to the size and connectedness of pore spaces, soil structure, and texture, and to the soil's composition (e.g. organic matter content, soil minerals) (Rezanezhad et al., 2016; Tuller and Or, 2005). Knowledge of the WRC is necessary to characterize the distribution and transport of water in soils, which are both required for hydrological modeling.

Several methods that spatially interpolate and/or extrapolate from point measurements are commonly applied. Most of the studies, that regionalized soil data, focus on either of three approaches (Herbst and Diekkru, 2006): regression approaches including multiple linear regression as well as machine learning algorithms (e.g. Hengl et al., 2017; Ließ et al., 2016), geostatistical approaches (e.g. Goulard and Voltz, 1993; Sinowski et al., 1997; Voltz and Goulard, 1994), and hybrid approaches (e.g. Haghverdi et al., 2015; Herbst and Diekkru, 2006).

In the case of the WRC, the regionalized variables are either single retention values of the curve (e.g. at field capacity and permanent wilting point, as in Haghverdi et al., 2015) or parameters of functions used to describe the WRC (e.g. in Yang et al., 2015), like the van Genuchten, Brooks-Corey or Campbell water retention functions (e.g.

* Corresponding author.
E-mail address: mareike.liess@ufz.de (M. Ließ).

https://doi.org/10.1016/j.geoderma.2017.12.002
Received 14 July 2017; Received in revised form 30 November 2017; Accepted 6 December 2017
Available online 19 December 2017

Khlosi et al., 2008; Too et al., 2014). An advantage of the latter is that the employed function - and its predicted parameter values - can easily be incorporated into simulation models (Wösten et al., 2001). Specific functions have been successful for certain parts of the WRC, certain soil types or regions of the world, but none of them is universal (Botula et al., 2014; Too et al., 2014). With the available information, the selection of the appropriate function is at present largely determined by convenience of the researcher (Too et al., 2014) and, therefore, represents itself a problem that is out of the scope of this study.

The objectives of this study are to present new water retention data of Páramo soils and to set up a model to regionalize point measurements of the WRC at common pF values at the catchment scale using the random forest algorithm (Breiman, 2001). In digital soil mapping it has been used for manifold applications regarding different soil properties (Grimm et al., 2008; Guo et al., 2015; Hengl et al., 2015; Ließ et al., 2016; Wiesmeier et al., 2011). Whereas, to our knowledge, random forest has not been used to spatially predict the water retention of soils, before. Thus, we take this opportunity to investigate its performance in this regard and to estimate the relative importance of predictor variables, which correlate with hydrological properties of soils (e.g. Thompson et al., 2012). We are not aware of the existence of any other studies that aimed to regionalize the WRC of Páramo soils, before.

2. Methods

2.1. Research area

The Quinuas Catchment (Fig. 1) is located in the western part of the Paute river basin and covers an area of approximately 93 km^2 comprising part of the Cajas National Park. Located at around 2.8°S and 79.2°W between 3000 and 4400 m.a.s.l. (Ließ, 2015), the catchment's conditions match that of the wet Páramo ecosystem (Arroyo et al., 2013; Hofstede et al., 2003).

Due to its location in an inner Andean valley close to the Equator and due to the relatively narrow transversal section of the Andes at these latitudes, the climate of the catchment is influenced by air masses coming from both the Pacific ocean and the Amazon basin (Buytaert et al., 2006b; Celleri et al., 2007). Mean annual temperature in the catchment varies between 5.3 and 8.7 °C without seasonality, while total solar radiation, wind speed, and rainfall vary seasonally (Carrillo-Rojas et al., 2016; Córdova et al., 2016). Mean annual precipitation ranges from 900 to 1600 mm between 2980 and 4100 m.a.s.l. (Crespo et al., 2011); year-round drizzle, accounts for 29% of the total annual rainfall amount (Padrón et al., 2015). Rainfall is characterized by a bimodal pattern with one early peak March/May, and the other in October (Celleri et al., 2007; Padrón et al., 2015); following the October rainfall peak, increased solar radiation in November leads to a relative decrease of humidity. Because of the humid and cold conditions of the Páramo, along with volcanic ash inputs from the Quaternary volcanic activity (Barberi et al., 1988; Buytaert et al., 2007), low density, porous soils rich in organic material developed across the Paute basin (Buytaert et al., 2007; Poulenard et al., 2003). The soils have a high water storage capacity and a high saturated hydraulic conductivity (Buytaert et al., 2006c). The prevailing vegetation in the catchment is tussock grass (*Calamagrostis* sp. and *Festuca sp.*), which is present in > 70% of the area and coexists with cushion plants (e.g. *Plantago sp., Valeriana* sp. and *Gentian sp.*), small forest patches of *Polylepis* sp. and *Gynoxis* sp. and low shrubs like *Weinnmannia sp.* (Carrillo-Rojas et al., 2016). The occurrence of cushion plants increases above 4000 m.a.s.l. (Sklenář and Jørgensen, 1999).

2.2. Dataset

2.2.1. Soil data

Undisturbed soil samples were collected with 100 cm^3 steel cores of 4 cm height during a field campaign in 2014. The sampling design, explained in detail by Ließ (2015), was based on the following tenets: 1) stratified random sampling according to landscape characteristics and 2) accessibility of the area and sampling costs, while following the concept that similar landscape positions carry similar soils with similar soil properties. We used the "QC-arLUS" sampling design among the four suggested designs in Ließ (2015) but sampled only two of the selected points in each landscape unit due to the time-consuming laboratory work in determining water retention curves. This resulted into 48 sampling locations. Samples were taken at 7 cm depth from the surface (steel core sample from 5 to 9 cm) to avoid the root felt.

θ at pF 0, 0.5, 1.5 and 2.5 was measured by placing the water

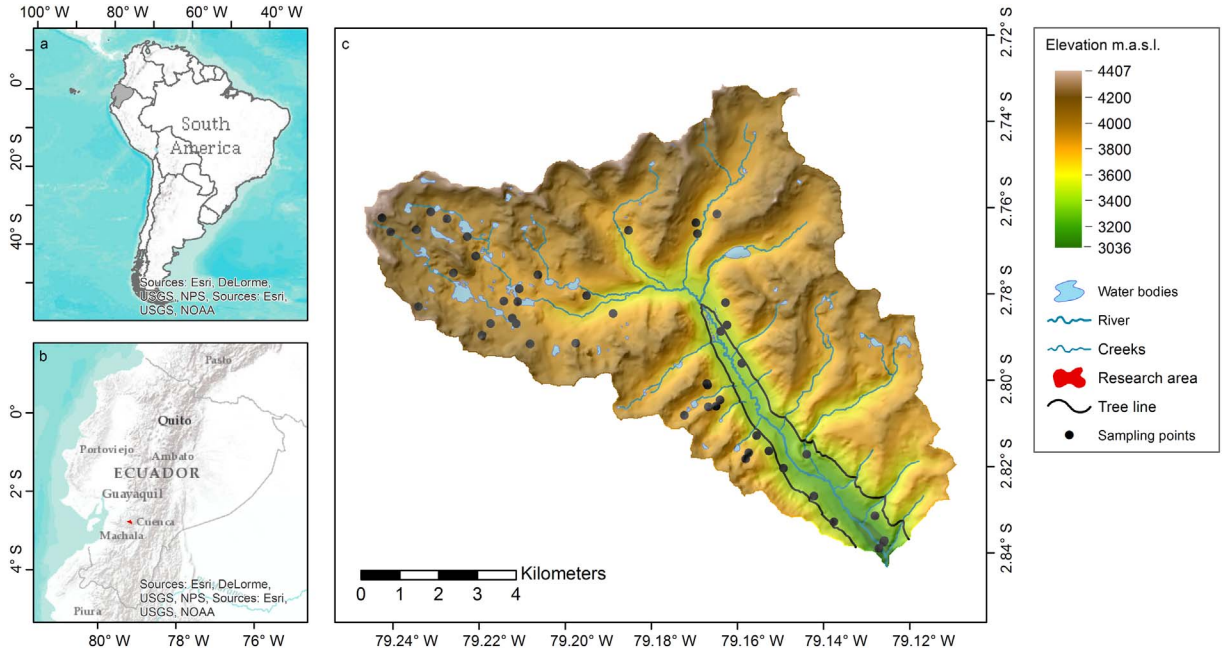

Fig. 1. Research area and sampling locations. a) Ecuador in South America, b) Research area within Ecuador, c) Location of sampling points within the research area (Overlaid hillshading with light source from North-West). Topographical data use with permission from the Ecuadorian Geographical Institute (2013, national base, scale 1:50.000).

Table 1
Predictors derived from the Landsat 8 image used as predictor variables. Physical interpretations are based on (Carlson and Ripley, 1997; USGS, 2013b; Wiegand et al., 1991).

ID	Predictor	Input	Physical interpretation
1	Reflectance 2	Band 2	Blue radiation; soil-vegetation contrast; deciduous -coniferous vegetation contrast
2	Reflectance 3	Band 3	Green radiation; vegetation
3	Reflectance 4	Band 4	Red radiation; vegetation
4	Reflectance 5	Band 5	NIR radiation; biomass content
5	Reflectance 6	Band 6	SWIR radiation; moisture content of soil and vegetation
6	Reflectance 7	Band 7	SWIR radiation, moisture content of soil and vegetation
7	Thermal 1	Band 10	TIRS radiation; soil moisture
8	Thermal 2	Band 11	TIRS radiation; soil moisture
9	NDVI	Reflectance 4 and 5	Plant characteristics, e.g. leaf area, vegetation condition, green leaf density, photosynthetically active biomass, water content
10	NDWI	Reflectance 5 and 6	
11	PVI	Reflectance 4 and 5	
12	TSAVI	Reflectance 4 and 5	

saturated steel core samples in a sandbox device and applying hanging water columns of increasing length (Durner et al., 2005; Veerman and Stolte, 1997). Data which showed unrealistic θ values, i.e. out of the range 0 to 1 were discarded, resulting in observations for 47 of the locations of at least triplicate measurements. We did not succeed in determining θ at pF values higher than 2.5 due to the high organic matter content of the samples.

2.2.2. Predictors

Model predictors were calculated from a Landsat 8 image and a digital elevation model (DEM), both with a 30 × 30 m resolution. The derived vegetation indices and terrain parameters (Table 1 and Table 2) were used as predictor variables to regionalize the point data. Vegetation indices and terrain parameters are common proxies for organic activity and topography (Thompson et al., 2012), which are among the factors of soil formation suggested by Jenny (1994).

The Landsat 8 image provided by the USGS (USGS, 2013a), was recorded on the 15th of September 2015. Among all the available Landsat 8 images (until July 2016) we finally selected the one with the least amount of clouds covering the catchment. The signal was transformed to reflectance or thermal signal according to the instructions from USGS (USGS, 2013b). Details about the physical meaning of the predictors and references for their calculation are shown in Table 1. Calculations were conducted in the software environment and programming language R (R Development Core Team, 2016). The Normalized Difference Vegetation Index (NDVI) (Rouse et al., 1973) and the Normalized Difference Water Index (NDWI) (Gao, 1996) were calculated using ratios between band differences as shown in Jackson et al. (2004). The Transformed Soil Adjusted Vegetation Index (TSAVI) and the Perpendicular Vegetation Index (PVI) were calculated with parameters derived from the 'Bare Soil Line' function (BSL) (Baret and Guyot, 1991; Fox et al., 2004) and bands 4 and 5 (red and NIR respectively). The BSL function estimates a line tangent to the lowest points of the roughly triangular distribution observed when the values of the NIR spectral band are plotted against the corresponding red band values of a given image. Pixels containing only bare soil lie at the base of the triangle, whose corners represent pixels containing the brightest (driest) and darkest (wettest) soils (Maas and Rajan, 2010). R's BSL function is available in the landsat package (Goslee, 2011). It uses the quantile method (Goslee, 2015) and user-defined upper and lower limits: in our case, *ulimit* = 0.999 and *llimit* = 0.013. The parameters used to calculate PVI and TSAVI are the intercept and the slope of this linear function.

Table 2
Terrain parameters used as predictor variables.

ID	Predictor	SAGA module	Definition	Physical interpretation
13	Altitude (DEM)	Multilevel B-spline interpolation	Elevation	Local climatic gradients, vegetation patterns, potential energy
14	Flow accumulation	Flow accumulation (flow tracing)	Area draining to catchment outlet	Runoff volume
15	SAGA - Topographic Wetness index	SAGA Wetness index	Estimate of water accumulation in an area	Soil moisture distribution
16	Aspect - easting	Morphometric features	Slope azimuth	Solar insolation evapotranspiration, vegetation distribution, and abundance
17	Aspect - northing			
18	Mass balance index	Mass balance index	Index based on slope magnitude, slope curvature and vertical distance to channel network	Net erosion/net deposition areas and related landforms
19	Plan curvature	Morphometric features	Contour curvature	Converging/diverging flow, soil water content
20	Profile curvature	Morphometric features	Slope profile curvature	Flow acceleration/erosion, deposition rate
21	Slope	Morphometric features	Surface gradient	Overland and subsurface flow velocity and runoff rate
22–26	Terrain surface convexity	Terrain surface convexity	Measure of surface upwards convexity	Low relief, lowland forms
27–31	Terrain surface texture	Terrain surface texture	Measure of surface roughness: accounts for the density of pits and peaks in an area	Types of lithology
32–36	Terrain ruggedness index	Terrain ruggedness index	Measure of terrain ruggedness/heterogeneity	Degree of site exposure
37	Valley Depth	Relative heights and slope positions	Vertical distance to a channel network base level	Relative gravity potential at a given position
38	Diffuse insolation	Potential incoming solar radiation	Diffuse insolation received from the sky's hemisphere	Evapotranspiration, plant growth
39	Direct insolation	Potential incoming solar radiation	Direct insolation received from the sun disk	Evapotranspiration, plant growth
40	Negative openness	Topographic openness	Measure of degree of enclosure of a location on an irregular surface	Concave reliefs, degree of exposure
41	Positive openness	Topographic openness	Measure of degree of dominance (positive) of a location on an irregular surface	Convex reliefs, geological/geomorphological features, degree of exposure

The digital elevation model (DEM) was derived from a 40 m distance contour lines shapefile provided by the Ecuadorian Instituto Geográfico Militar (IGM 2012, national cartographic base of scale 1:50,000) using the multilevel B-spline algorithm (Lee et al., 1997) and preprocessed by filling the sinks. Terrain parameters were derived from the DEM using the SAGA GIS software (System for Automated Geoscientific Analyses) (Conrad et al., 2015). The set of DEM-derived parameters included in this study are related to θ, local climatic gradients and vegetation patterns (Table 2). Definitions and physical interpretations are based on Böhner and Antonic (2009), Böhner and Selige (2006), Conrad (2012), Gruber and Peckham (2009), Iwahashi and Pike (2007), Ließ et al. (2016), Mercado et al. (2009), Möller et al. (2008), Moore et al. (1991), Olaya (2009), Oliveira et al. (2010), Riley et al. (1999), and Yokoyama et al. (2002). Elevation, slope, aspect, plan and profile curvatures and flow accumulation are known as primary parameters, being directly calculated from the DEM (Olaya, 2009). The rest, known as secondary parameters, were computed from two or more primary parameters. Terrain parameters are useful for soil regionalization in environments where topography strongly influences soil development (McKenzie et al., 2000). The parameters Terrain Surface Convexity, Terrain Surface Texture and Terrain Ruggedness Index (Table 2, ID 22–36) were calculated with search radii of 5, 10, 15, 20 and 30 cells. We checked for distribution function, outliers, collinearity and variance of the response and predictor variables by means of graphical methods (Zuur et al., 2010) to better decide on our modeling approach and understand its limitations for our data. Our major concerns were due to collinearity among predictor variables (Fig. 2), extreme observations in both predictor and response variables and the small size of our dataset. As a consequence, special importance has to be given to predictor selection and the model training and assessment procedure, as will be shown in Chapters 2.4 and 2.5.

2.3. Random forest

Random forest models (Breiman, 2001) are ensembles of decision trees. In each tree, the dataset is recursively partitioned into increasingly homogeneous subsets regarding the response variable based on the improvement of the residual sum of squares (RSS). All predictors are tested at each split to decide which predictor and which value of the predictor subdivides the dataset best. Random forest is regarded as a non-linear, non-parametric method that is able to handle small-n-large-p problems (Díaz-Uriarte and Alvarez de Andrés, 2006; Grömping, 2009; Strobl et al., 2007; Touw et al., 2013), seldom overfits (Breiman, 2001; Hastie et al., 2009a, 2009b) and is relatively robust to outliers and noise (Díaz-Uriarte and Alvarez de Andrés, 2006).

Random Forest differs from decision tree models in growing many trees instead of a single decision tree and averaging the results. The number of trees is referred to as *ntree*. Model strength is achieved by two means: (1) by varying the input dataset, i.e. taking a bootstrapped sample of size equal to two-thirds of the full dataset for training each tree, and (2) by varying the tree model structure, i.e. selecting a random predictor subset from all predictors at each tree node. The size of this predictor subset - here referred to as *mtry* - remains the same for the whole forest. The observations left out from model training, called “out of bag” (OOB), are used to estimate model accuracy (Breiman, 2001). For regression models, the prediction error is returned as mean squared error (MSE) (Grömping, 2009). We used the train() function of the R package **caret** for model training and employed as a dependency the implementation of randomForest() by Liaw and Wiener (2002). The set.seed(120) function was used in order to obtain reproducible results.

Two variable importance metrics are returned by random forest: (1) the increased node purity, which attributes the improvement in the split-criterion RSS as importance measure to the splitting variable (Hastie et al., 2009a, 2009b), and (2) the permutation importance (also known as MSE reduction) (Grömping, 2009). Variable importance is commonly used to aid in the interpretation of the dataset by uncovering

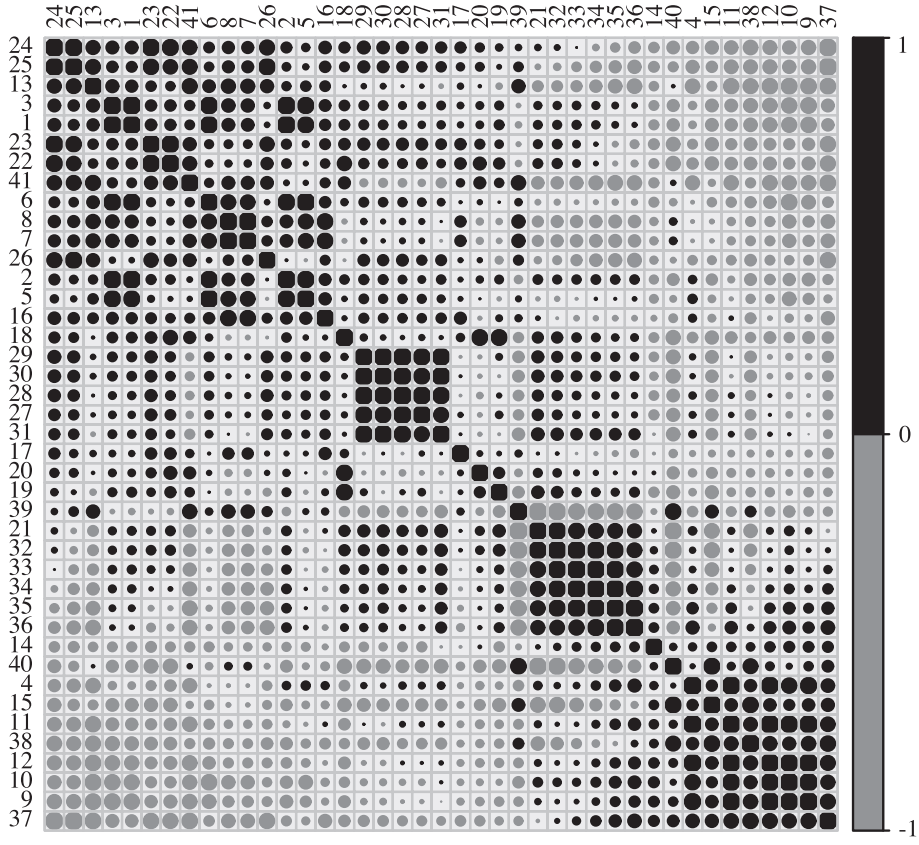

Fig. 2. Correlation plot of predictors. Black dots correspond to positive values and grey dots to negative values of Pearson's correlation coefficient. The size of the dots represents the magnitude of the coefficient.

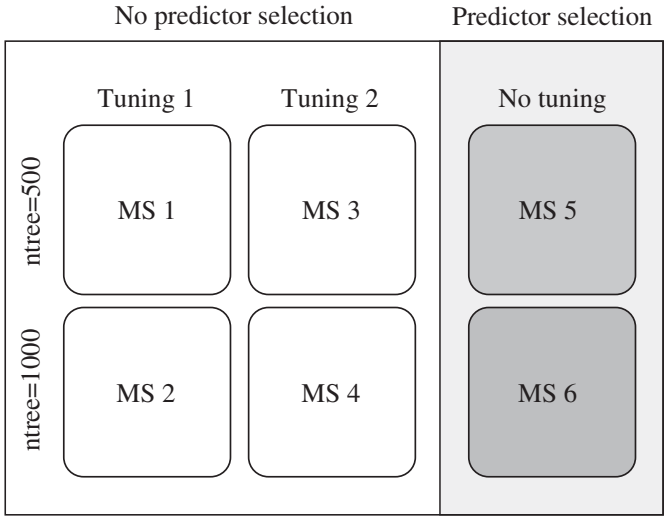

Fig. 3. The six model setups (MS) including two *ntree* values, mtry tuning, and predictor selection.

interactions between predictor variables (Breiman, 2001; McKinney et al., 2006; Touw et al., 2013; Wright et al., 2016), identifying important predictor variables and as a filter to remove non-informative predictor variables (Díaz-Uriarte and Alvarez de Andrés, 2006).

2.4. Model setup

Especially in the case of small datasets, the influence of model parameter tuning and predictor selection on model performance in machine learning seems worthwhile investigating. To understand the effect of predictor selection and *mtry* tuning while using different *ntree* values, we tested six model setups. In Fig. 3 tuning refers to the process of selecting the best *mtry*. Predictor selection was done with "Recursive Feature Elimination" (RFE). Both tuning and RFE, are explained in detail below.

2.4.1. Model parameter selection

The idea behind tuning is to find the value for the one or several parameters that produce the minimum in the test error curve (Hastie et al., 2009a, 2009b). According to Breiman (2001), reducing the correlation between any two trees in the forest and increasing the strength of individual trees decrease the prediction error. Reducing *mtry* reduces both correlation and strength. Hence, *mtry* is a sensitive tuning parameter to improve model performance. The default *mtry* value of randomForest() (Liaw and Wiener, 2002) relates to the number of predictors (p) and is $p/3$. However, as with any machine learning algorithm, the optimal parameter value depends on the data (Hastie et al., 2009a, 2009b). In the case of random forest, this is particularly true for datasets with correlated predictor variables, for which several *mtry* values should be considered (Strobl et al., 2008). When using a big number of predictors- several of which might be noise - the optimal *mtry* would be higher than the default (Breiman, 2003; Genuer et al., 2010). Breiman (2003) recommends trying as well with half of the default value and twice of the default value. We used the *tuneGrid* argument in the train() function of **caret** to search for the optimal *mtry* within a predefined range of *mtry* values. This is based on the suggestions that (1) the optimal *mtry* should be around $p/3$ (Breiman, 2001; Liaw and Wiener, 2002), and (2) if several predictor variables are noise, the optimal *mtry* would be higher than $p/3$ (Breiman, 2003; Genuer et al., 2010). Accordingly, a one-dimensional grid search was implemented with two different parameter ranges referred to as Tuning 1 and Tuning 2:

Tuning 1: $mtry = \frac{p}{3} - n, \ldots, \frac{p}{3}, \ldots, \frac{p}{3} + n$; *for* $n = 1, 2, 3, 4, 5$.

Tuning 2: $mtry = \frac{2p}{3} - n, \ldots, \frac{2p}{3}, \ldots, \frac{2p}{3} + n$; *for* $n = 1, 2, 3, 4, 5$

According to Breiman (2003), random forest can grow as many trees as desired and it will not overfit. In contrast, Hastie et al. (2009a, 2009b) suggest that too many trees would result in too rich a model, which could increase the variance. Here, we set *ntree* to 500 or 1000 instead of tuning it.

2.4.2. Predictor selection

The purpose of predictor selection for supervised learning methods is to increase model accuracy or to reduce model complexity (Kuhn and Johnson, 2013). Kuhn and Johnson (2013) showed that including non-informative predictors in a random forest model leads to moderate degradation in performance. In predicting soil organic carbon stocks with random forest, Ließ et al. (2016) and Xiong et al. (2014) showed different degrees of improvement in model performance using different predictor selection algorithms. Using the Boruta selection algorithm, Ließ et al. (2014) showed that the prediction of one out of three soil parameters improved. Because of the potential influence of predictor selection in model performance, we tested two approaches: (1) without predictor selection (model setups 1 to 4, Fig. 3) and (2) with predictor selection (model setups 5 and 6, Fig. 3), using the RFE algorithm (Guyon et al., 2002; Kuhn and Johnson, 2013).

RFE is a backward predictor selection algorithm that uses the permutation importance measure of random forest as a ranking criterion (Kuhn and Johnson, 2013). The permutation importance is computed by measuring how the OOB prediction error changes when the model loses access to the true values of a given predictor variable, i.e. when the association between the response variable and a predictor variable is broken. To simulate this "induced independence" between the response variable and a predictor variable, the OOB values of the predictor variable are randomly scrambled, i.e. permuted. To get the importance measure, the MSE is calculated before and after permutation of the observations, and the difference is averaged over all trees. These differences are often expressed as percent of the maximum (Grömping, 2009; Hastie et al., 2009a, 2009b). According to Gregorutti et al. (2016) RFE reduces the effect of correlated predictors on the importance measures of random forest.

The rfe() function implemented in **caret** (Kuhn, 2008, Kuhn and Johnson, 2013) incorporates a resampling step that deals with concerns due to selection bias (Ambroise and McLachlan, 2002) and improper use of resampling (Kuhn and Johnson, 2013). In the case of the models trained with RFE, *mtry* is set to $p/3$ and varies only due to the number of predictors selected by the model. The search for the best number of predictors is controlled by a search interval, with the argument *sizes* of the rfe() function, whose default is 2^n *with* $n = 2, 3, 4$. We expanded the search interval to *sizes* = 1:10, 10, 15, 20, 25, 30, 36.

2.5. Modeling procedure and model assessment

The man aim of modeling is to form a robust model. Statistical learning algorithms are highly adaptable, which means that they can very easily overemphasize data structures that are not reproducible. Many algorithms include sensible tuning parameters which further strengthens the algorithms' flexibility. Accordingly, overfitting is an aspect which has thoroughly been discussed in the literature (e.g. Hastie et al., 2009a, 2009b; Hawkins, 2004). To avoid overfitting and obtain robust model performance estimates, usually, some sort of resampling is used; models are trained and evaluated on various data subsets. The training set is used for model training and tuning, the test set is used to evaluate the model's predictive performance. Finally, for model tuning it is a good choice to also use resampling in order to obtain robust tuning parameter values (e.g. Hastie et al., 2009a, 2009b; Kuhn and Johnson, 2013). The same applies for predictor selection.

Along the validation steps that we followed, we differentiate between *model setups, models and model runs*. Each *model setup* is

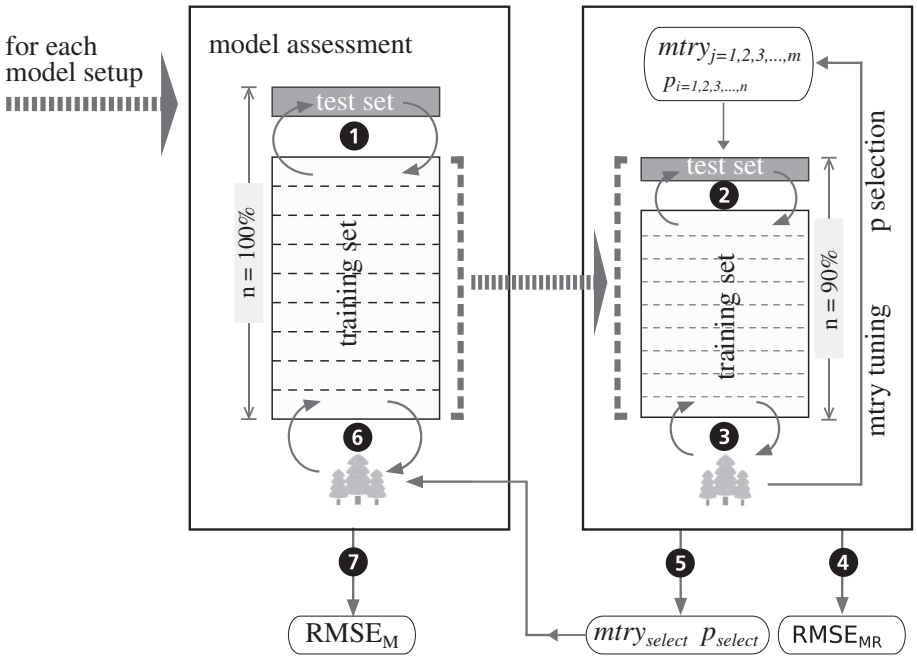

Fig. 4. Model selection and validation procedure. The box on the left represents the model validation procedure. The box on the right represents the model tuning and predictor selection procedure. The tree icons at the bottom of each dataset represent the random forest algorithm, with the circular arrows meaning the interaction between a model and the dataset, for model training, test set prediction or both. The black circles correspond to the modeling steps.

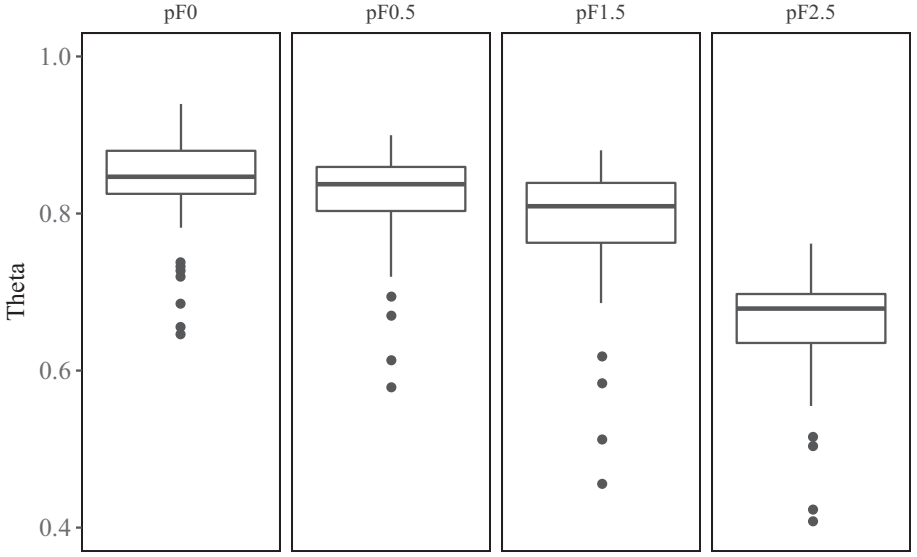

Fig. 5. Boxplots of volumetric water content (θ) at different pF values.

characterized by a combination of *mtry* tuning, *ntree* setting, and/or predictor selection (Fig. 3). Model setup validation is done by 5 times repeated 10-fold cross-validation (5R10CV). Accordingly, 50 training sets were created from the dataset, a *model* is trained with each of them and validated with the remaining test set data (left box, Fig. 4) resulting in 50 models and 50 predictions respectively. The median of these 50 predictions will be displayed on a map, the interquartile range (IQR) will be used as a site-specific performance estimate. Each of the *models* is trained with a certain *mtry* and *p* determined by tuning and predictor selection, respectively. Parameter tuning and predictor selection are also done by 5R10CV (right box, Fig. 4).

The complete procedure of model tuning, predictor selection, and model setup validation involves a number of steps illustrated in Fig. 4. For each model setup, the full dataset was partitioned in step 1, using 5R10CV via the function createMultiFolds() of **caret**. This resulted in 50 training and test sets of n = 90% and n = 10% of the full dataset, respectively. Each of the thereby created training sets was partitioned into 50 "sub" training and test sets to conduct predictor selection and parameter tuning via 5R10CV in step 2 This was implemented using the function trainControl() or rfeControl() of **caret**. In step 3, several *model runs* were conducted each of which tried a different $mtry_j$ or p_i, according to the tuning or predictor selection procedure defined by the model setup (Fig. 3). The performance of each *model run* was assessed by calculating the 50 RMSE ($RMSE_{MR}$) based on the "sub" test sets in step 4. The *model run* with the lowest median $RMSE_{MR}$ was identified and its *mtry* ($mtry_{select}$) or *p* (p_{select}) was selected in step 5 to define a *model*. In step 6, each *model* was trained based on the particular training set with the specified $mtry_{select}$ and p_{select}. The corresponding test set RMSE of each model ($RMSE_M$) was calculated and stored in step 7 (Casella et al., 2006; Hastie et al., 2009a, 2009b), the actual model assessment. Altogether, this means that for each pF six *model setups* were tested resulting in 50 *models* each.

3. Results and discussion

3.1. Water retention data

The θ values measured at different pF values ranged from 0.36 to

0.94. With increasing pF, the median θ decreased from 0.85 at pF 0 to 0.68 at pF 2.5 and its distribution became more skewed towards low values (Fig. 5). Some of the samples show comparatively much lower θ values at almost all pF values and even exceed the boxplots' fences in Fig. 5. However, compared to the other samples which have organic carbon contents and bulk densities which are typical for organic soils, their bulk densities are significantly higher (0.81 to 0.52) and their carbon contents are significantly lower (8.8% to 24%) (data not included) indicating that these samples correspond to soils with a higher proportion of mineral components. These observations were therefore not treated as outliers but kept in the dataset. However, the dataset also includes samples which have comparatively low values in both, organic carbon content and bulk density, and we assume that these are soils with andic properties. Altogether, data mining has shown (results not included) that particularly for these Páramos soils, the distinction between organic and mineral soils due to the organic carbon content reported in textbooks and field guides for soil survey (e.g. Blume et al., 2010; FAO, 2006) subdivides the soil continuum by mere definition.

Most of the observed water retention values are similar to data from other Páramo areas in south Ecuador (Buytaert et al., 2005) and data from other peat soils (Boelter, 1966; Gnatowski et al., 2010; Schwarzel et al., 2002; Schwärzel et al., 2006), but higher than values obtained for Andosols (Arnalds et al., 2007) and mineral soils from Europe (Hewelke et al., 2015). According to Buytaert et al. (2005, 2006a, 2006b, 2006c), histic Andosols from Páramo landscapes have an open and porous structure, which is stronger than that of peat soils. It is not clear though whether the water retention might be controlled by similar factors. In the case of peat soils, the most important controlling factors are the cell structure of the plant residues, i.e. the percentage of plant tissue (Weiss et al., 1998) and the degree of organic matter decomposition (Boelter, 1966, 1969; Rezanezhad et al., 2016). Minaya et al. (2016) studied the plant residues of Páramo soils in northern Ecuador and their relation to the degree of decomposition and altitude. In their study, they identified a higher decomposition degree related to acualescent rosettes and cushions than tussock species. Because similar plants are observed in the Quinuas catchment, we hypothesize that the composition of plant residues might play a role as well in the soils we studied.

Soils of Páramo ecosystems in Ecuador have been referred to as volcanic ash soils with high organic carbon contents and hydric Andosols (Buytaert et al., 2006c, 2007) and thus, similarities regarding water holding mechanisms are expected. Water retention could be due to absorption/adsorption by short-range ordered minerals or complexation with cationic metals. It is expected that the histic Andosols widespread in Páramo landscapes from southern Ecuador favor the second mechanism (Buytaert et al., 2005). Compared to peat soils and Andosols, there is a lack of understanding about the relative contribution and interplay of andic and organic materials regarding the water retention at different pF values in Páramo ecosystems.

3.2. Model tuning and predictor selection

The outputs of the model selection procedure were 50 models for each model setup and pF, each of which was defined by a certain *ntree*, *mtry* and *p* (step 5, Fig. 4). Fig. 6 shows the boxplots of the selected $mtry_{select}$ values after tuning. Please remember that model setup 5 and 6 do not involve any *mtry* tuning (Fig. 3). The observed variation of *mtry* in these two model setups is caused by predictor selection, because $mtry_{select} = \frac{p_{select}}{3}$.

For model setups 1–4, at pF 0, 0.5 and 1.5, *mtry* values ($mtry_{select}$) close to the maximum of their search intervals were selected: close to 19 predictors in the case of model setups 1 and 2, 32 in the case of model setups 3 and 4, suggesting that even higher *mtry* values might have had to be checked. At pF 2.5 the situation is different: for model setups 1 and 2 the selected *mtry* values ($mtry_{select}$) are again close to the maximum, but for model setups 3 and 4, the selected *mtry* values ($mtry_{select}$) are close to the minimum. This suggests that the smallest median $RMSE_{MR}$ might have been reached at some point between the *mtry* search intervals of Tuning 1 and Tuning 2. Regarding the spread of the selected *mtry* ($mtry_{select}$) within the same tuning procedure, in several cases, the model setup with *ntree* = 500 showed a larger variability than with *ntree* = 1000. Within the same *ntree*, the variability of $mtry_{select}$ values was generally larger (all pF) for model setups tuned with higher *mtry* values (Tuning 2) compared to those that were tuned with lower *mtry* values (Tuning 1). As for the number of selected predictors (model setups 5 and 6), Fig. 6 shows a decreasing number of p_{select} (mtry*3) with increasing pF. Few predictors provide sufficient information to predict water retention at higher pF values. A low spread of selected *p* (p_{select}) is observed at all pF values, except for pF 0.5.

Fig. 7 and Fig. 8 show the $RMSE_{MR}$ (Fig. 4, step 4) related to the tuning of *mtry* and the predictor selection in the form of test error curves. In Fig. 7 the values of the light grey lines corresponding to each $mtry_j$ value are the 50 $RMSE_{MR}$ for each model run. The dark grey and black lines correspond to the mean and median $RMSE_{MR}$. The *mtry* value corresponding to the lowest median $RMSE_{MR}$ ($mtry_{select}$) was selected to define a model. In Fig. 8 the values of the light grey lines corresponding to each p_i value are the 50 $RMSE_{MR}$ for each model run. The dark grey and black lines correspond to the mean and median $RMSE_{MR}$. The number of predictors, *p,* corresponding to the lowest median $RMSE_{MR}$ (p_{select}) was selected to define a model. The higher *mtry* values of the Tuning 2 search interval applied for model setups 3 and 4 lead to lower $RMSE_{MR}$ values (Fig. 7). This can be observed for the models adapted for the water retention at pF 0, 0.5 and 1.5, while *mtry* tuning does not seem to have a large impact on the models predicting water retention at pF 2.5. The tendency towards a $mtry_{select}$ value at the higher end of the Tuning 1 search interval and the lower end of the Tuning 2 search interval for the models corresponding to this pF value could also be observed in Fig. 6. The rather steep negative slope of the $RMSE_{MR}$ trend curves for model setups 1 and 2 compared to model setups 3 and 4 (pF 0.0, 0.5, 1.5) indicate a comparatively higher *mtry* tuning importance for the Tuning 1 search interval which surrounds the *mtry* default value of *p*/3 (Fig. 7). The error curves of the mean and median $RMSE_{MR}$ for the predictor selection procedure (Fig. 8) do not show much difference for model setups 5 and 6 indicating that the number of trees does not impact predictor selection. While the error curves of the mean and median $RMSE_{MR}$ show well-defined minimum peaks for pF 1.5 and 2.5 at a low number of predictors and at least an initial rapid decrease for pF 0.5, the error curves of the mean and median $RMSE_{MR}$ for pF 0.0 indicate the necessity of a high number of predictors which can also be observed from Fig. 8.

Fig. 7 and Fig. 8 report the error scenarios that happened during the cross-validated model tuning and predictor selection procedure. For studies that don't use resampling, the tuning or predictor selection procedure could be represented by any of the light grey curves. The representation of the many possible realities regarding tuning and predictor selection is seldom reported though. Given the variability between these curves, whether a study uses resampling or not - especially when dealing with small datasets - should be taken into account when interpreting the uncertainty of the predictions.

3.3. Model assessment and model setup selection

Predictions overestimate θ at low values and underestimate it at high values, which results in a linear trend of the residuals (Fig. 9). The largest positive residuals account for c. 20% of the maximum θ values for the models for all four pF values. The size of the largest negative residuals is influenced by those few rather low θ values typical for mineral soils (Fig. 5). Due to their low number the initially observed "outliers" cannot be separated from the other samples within the random forest recursive partitioning procedure. This is owed to the fact that the minimum amount of samples per tree node was set to 5. And reducing this number would have reduced model robustness. This effect increases with increasing pF values. It indicates that the adapted models

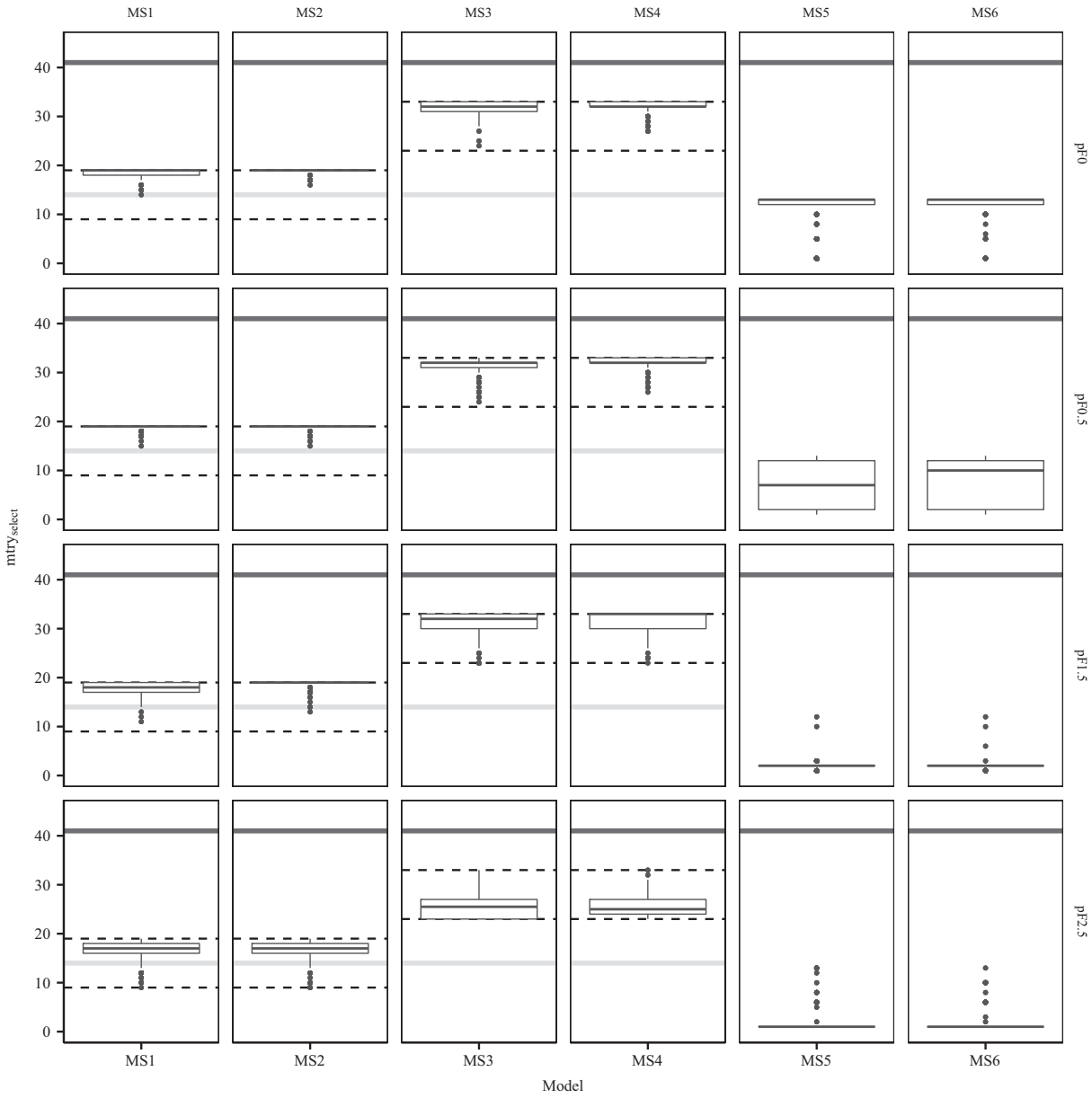

Fig. 6. Boxplots of the 50 selected mtry values ($mtry_{select}$) for each model setup (MS) and pF value. The lines serve as references. The solid dark grey lines correspond to the total number of predictors ($p = 41$), the solid light-grey lines correspond to p/3. The dashed lines are the upper and lower limits of the mtry search interval. For MS5 and MS6, the observed mtry variation is due to predictor selection.

cannot well predict water retention at pF 1.5 and 2.5 for mineral soils within the area and model performance would improve if those mineral soil samples were treated as outliers and removed from the dataset. The small size of the dataset, containing a few extreme observations was too small to capture landscape complexity. This was expected, but could not be avoided due to the limited available lab capacities. A similar pattern of residuals is reported in other studies concerning the regionalization of soil properties with random forest (Hengl et al., 2017; Mulder et al., 2016; Shangguan et al., 2016; Xu et al., 2016) and has been attributed to small datasets and underrepresented values of the target variable (Mulder et al., 2016) or to the prediction algorithm of random forest, which computes the unweighted average of the collection of trees (Xu et al., 2016). This creates results biased towards the sample mean, and, consequently, under/overestimation of large/small values of the target variable. The dataset, which was constrained in size and distribution due to site accessibility, sampling costs and lab capacities, sets our study apart from others like Haghverdi et al. (2015), Herbst and Diekkru (2006), Saito et al. (2008) and Voltz and Goulard (1994) who predicted soil water retention in gently sloping terrains and used regular and/or dense sampling designs, which allowed for the use of geostatistical and "hybrid" regression-interpolation approaches.

Fig. 10 summarizes the prediction errors ($RMSE_M$) of the 50 models for each model setup and pF. The model with the best predictive performance for each pF due to the lowest median $RMSE_M$ is marked in grey. With increasing pF a slight increase of the median $RMSE_M$ is observed: The lowest median $RMSE_M$ is 0.040 at pF 0 (MS2) and increases to 0.061 at pF 2.5 (MS4). The range of variation of $RMSE_M$ also increases with increasing pF, observable from the bigger size of the IQR, the larger extent of the fences and the increase of extreme values. Altogether, model setup performance was similar in terms of median $RMSE_M$ and range of variation. Slight differences of the median $RMSE_M$ and the $RMSE_M$ inter-quartile range (IQR) are nevertheless observed, most notably at pF 2.5: the IQR of $RMSE_M$ values differs among model setups that were built with different *mtry* tuning ranges, e.g. compare

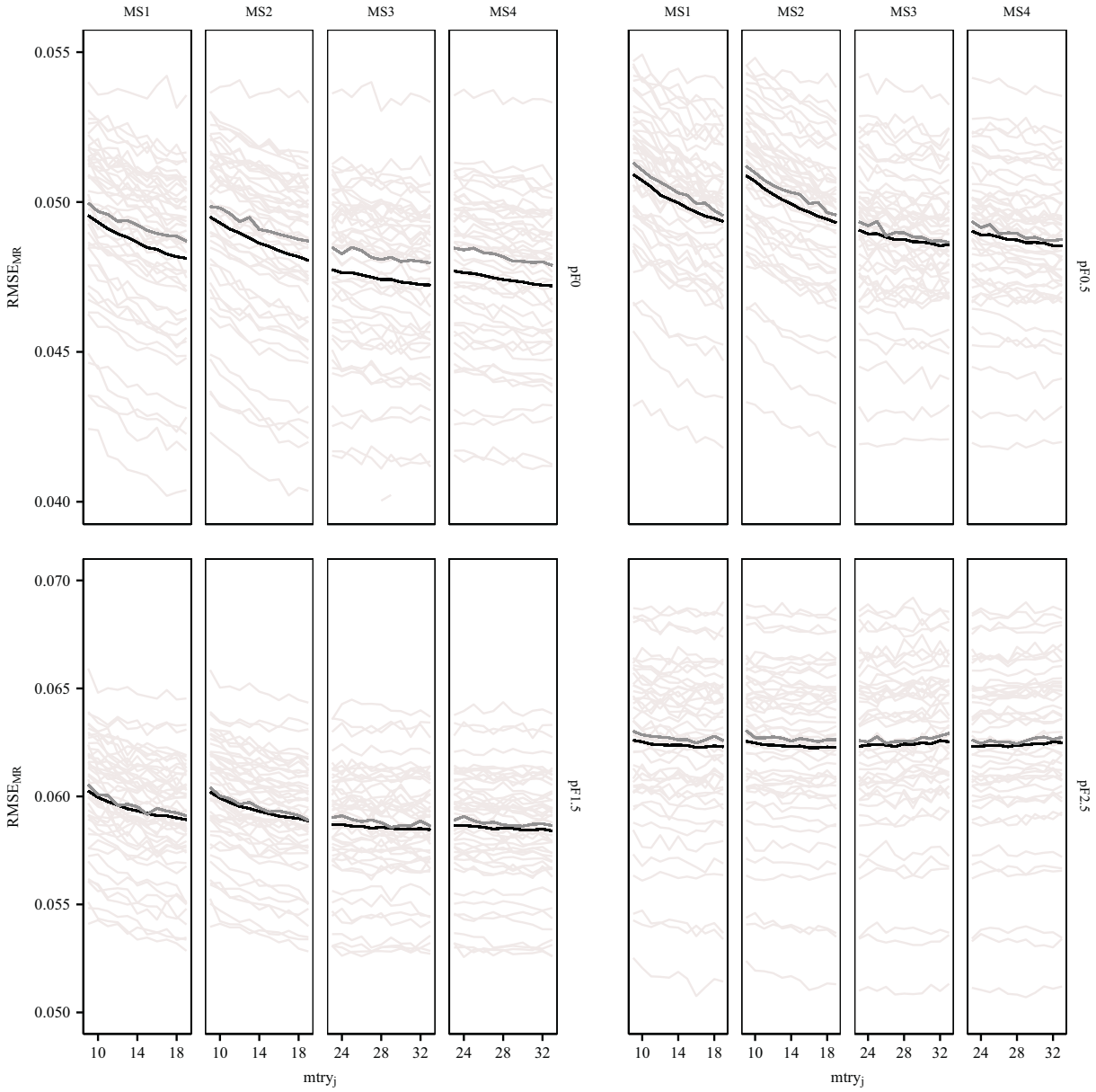

Fig. 7. Test error curves of mtry tuning for model setups (MS) 1–4. The values of the light grey lines corresponding to each $mtry_i$ value are the 50 $RMSE_{MR}$ for each model run. The dark grey and black lines correspond to the mean and median $RMSE_{MR}$. The mtry value corresponding to the lowest median $RMSE_{MR}$ was selected to define a model.

model setups 2 and 4, as well as among model setups with the same *mtry* tuning range but different *ntree*, e.g. model setups 1 and 2. Regarding the median $RMSE_M$, model setups 5 and 6 have lower values than the rest.

Fig. 7 and Fig. 8 show that *mtry* tuning and predictor selection improve model performance. A certain number of predictors are necessary to form a good model whereas adding further predictors beyond this point does not lead to further improvement. In contrast to this, predictor selection results for pF 0 (Figure 6and 8) indicate that the predictive performance (Fig. 10) could have been further improved by more predictors. And waiving predictor selection in this particular case might have led to an even better predictive performance. In all the other cases, it will surely help when interpreting the model results within the landscape context as it leads to less complex models. *mtry* seems to be a more sensitive parameter (Fig. 7). However, improvement in $RMSE_{MR}$ due to the tuning procedure was neglectable small. Other studies on tropical mountain soil-landscapes in Ecuador support the finding that the impact of tuning procedures on the prediction error is dataset dependent (Hitziger and Ließ, 2014; Ließ et al., 2014, 2016).

The variability within this small dataset is rather high. As a consequence, model setup comparison due to the tuning $RMSE_{MR}$ and the predictive performance ($RMSE_M$) does not necessarily go in line with one another. Regarding *mtry* tuning for pF 0, 0.5 and 1.5, model setups 3 and 4 lead to lower $RMSE_{MR}$ values compared to model setups 1 and 2 (Fig. 7). Accordingly, the model setup with the best predictive power in the case of pF 0.5 and pF 1.5 is model setup 3 (Fig. 10). On the contrary, for pF 0 model setup 2 was selected. For pF 2.5 the best model setups involved predictor selection instead of *mtry* tuning; the median number of selected predictors was 5. Relating the highlighted box-plots shown in Fig. 10 with the model setup characteristics shown in Fig. 3 reveals that, at all pF values but pF 0, the model setups with the best predictive performance were trained with 500 rather than 1000 trees indicating that the number of trees is, in fact, a model parameter which needs to be selected with care.

3.4. Spatial prediction

Fig. 11 shows the spatial predictions of the median (top row) and

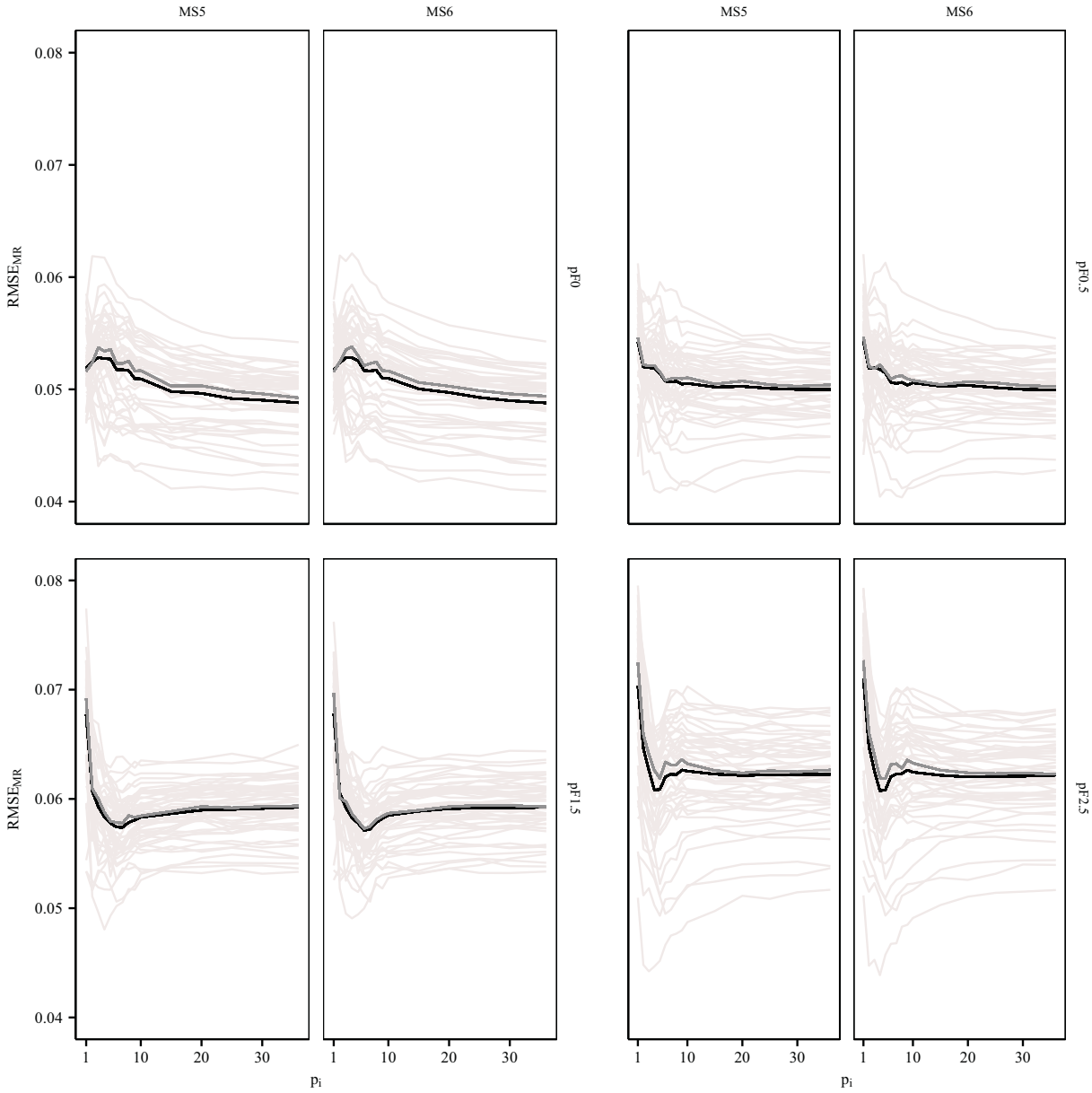

Fig. 8. Test error curves of predictor selection for model setups (MS) 5 and 6. The values of the light grey lines corresponding to each p_i value are the 50 $RMSE_{MR}$ for each model run. The dark grey and black lines correspond to the mean and median $RMSE_{MR}$. The p_i value corresponding to the lowest median $RMSE_{MR}$ was selected to define a model.

$-\log_{10}$(IQR) (lower row) θ values at the different pF values. The applied color scale of the map legend is based on parameter quantiles. Because the lowest and highest IQR differ by almost two orders of magnitude, and most of the values are in the middle range, the $\log_{10}$ transformation was used to color-differentiate the IQR distribution.

At pF 0 the highest θ values (0.85 to 0.9) are distributed as narrow stripes in tributary valleys and on the north-eastern exposed slopes of the Quinuas River valley. Values of the next lower quantile (0.83 to 0.85) are located on the summits and upper slopes. At pF 0.5 and 1.5 the three highest quantiles follow a similar distribution, but the areas with the highest median θ values become narrower on the upper tributary valleys and almost disappear downstream. At pF 2.5 most median θ values range between 0.67 and 0.7, and along the ridges, median θ is around 0.65. Finally, at all pF values, lowest θ values are observed in the lower part of the Quinuas river valley.

High IQRs correspond to low θ values which clearly shows the impact of the initially discussed extreme values (Fig. 5) corresponding to mineral soils. Areas with the lowest θ values include both, mineral and organic soils which could not be separated by the random forest models and are, therefore, assigned to the same area resulting into high model residuals and a high IQR in the predictions. At all pF values, the most uncertain predictions are located in the Quinuas valley. For pF 0 and 0.5, moderately uncertain predictions are located on the upper slopes and summits, whereas at pF 1.5 and 2.5 moderately to highly uncertain predictions are observed as broad stripes on the summits and slopes.

The most important predictors associated with the selected model setup for each pF, which may explain the spatial patterns of water retention (Fig. 11) are shown in Fig. 12. Importance scores which account for < 5% of the median score of the most important predictor (around 0.002) are not displayed. This resulted in six predictors for pF 0 to 1.5 and five predictors for pF 2.5. The order of importance of the top four predictors, altitude (ID 13), NDVI (ID 9), TSAVI (ID 12) and NDWI (ID10) remains the same for pF 0 to 1.5. Among the fifth and sixth most important predictors are reflectance 4 (ID 3), PVI (ID 11) and positive openness (ID 41). The latter distinguishes ridges from valley structures. We found that the importance of altitude (ID13) decreased relative to

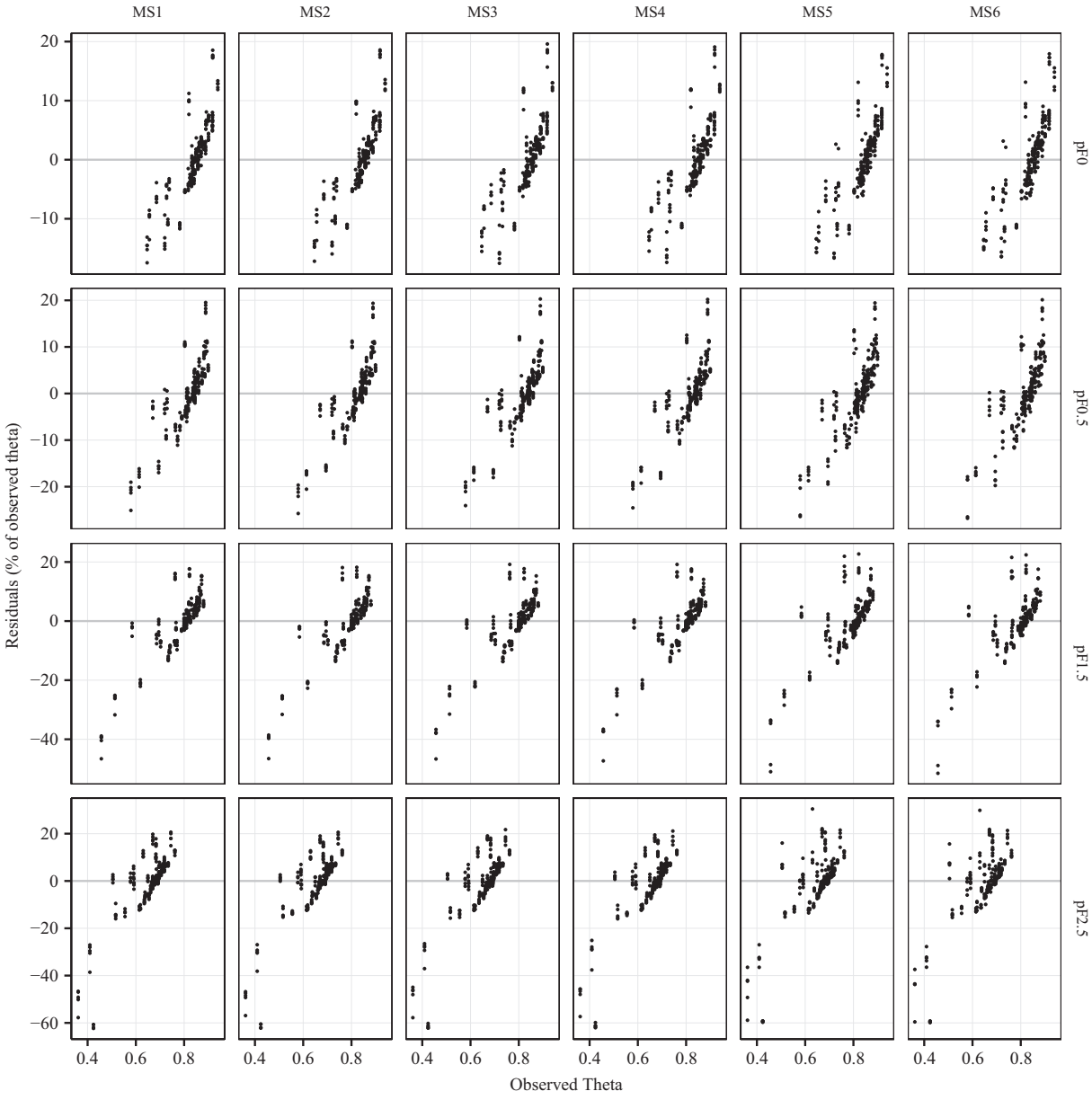

Fig. 9. Scatterplot of residuals versus measured water content (θ). For each observation, there were 5 predictions due to the 5 repetitions of the 10-fold cross-validation.

the importance of vegetation indices with increasing pF, to the point that, at pF 2.5, the order of importance reverses: here the top four predictors are vegetation indices and the least important is altitude. Geroy et al. (2011), Pachepsky et al. (2001) and Seibert et al. (2007) among others found topographic features useful in predicting soil water retention in mineral soils. Altitude as well as positive openness and of course the vegetation indices in the here studied Páramo landscape refer to the relatively higher importance of the spatial vegetation pattern in our models which can be attributed to the high organic carbon content of the soils, so that the type of the decaying plant material has a high impact on water retention in Páramo soils as was also suggested by Suárez et al. (2013). Pine forest plantations and pastures on the floodplains below the tree line (3440 m.a.s.l.) of the area indicate soil cultivation. Fluvial deposits and/or erosion of the organic topsoil together with soil cultivation might have contributed to the higher mineral soil content and the resulting lower water retention in this part of the catchment.

Landsat images representing vegetation type and soil exposure vary in time. The same applies for the vegetation indices calculated from them. Accordingly, a more robust approach would be to use a time series of satellite images. However, due to the frequent cloud cover and the temporal resolution of Landsat 8, this was not possible. Sentinel 2 data provide a chance to improve this situation (Paloscia et al., 2013). We tested soil bulk density, sessile drop contact angle data, and soil organic carbon content as additional predictor variables without improvement of the prediction error (data not shown). Future work should aim to find area-wide proxies of soil chemistry regarding mineralogy, organo-metallic complexes and degree of organic matter decomposition. In this regard, satellite images with higher spectral resolution might be useful, e.g. as shown in Garfagnoli et al. (2013) and Steinberg et al. (2016). Alternatively, using proxies in the form of soil point measurements require regionalization, which pose some difficulties because 1) correlation lengths of soil hydraulic properties and soil physicochemical properties differ at the catchment scale (van der Keur and Iversen, 2006), and 2) it would require an extensive and concomitantly expensive sampling campaign to allow for interpolation approaches. Nevertheless, soil point measurements could be used to increase the database of water retention values in Páramo soilscapes

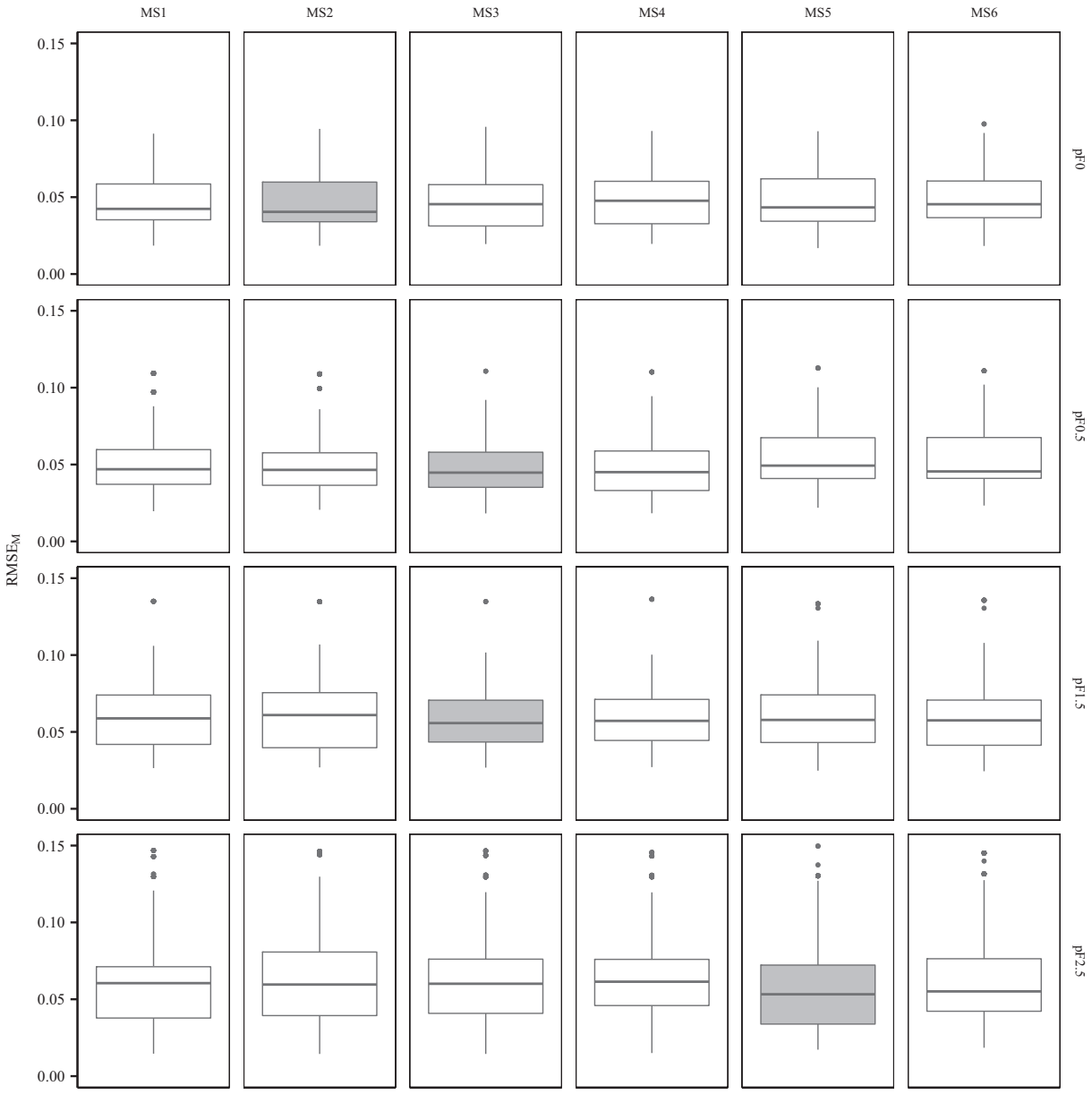

Fig. 10. Model validation results. Boxplots of the 50 $RMSE_M$ for each model setup (MS) and pF value. The boxplot with the lowest median $RMSE_M$ is colored grey.

using pedotransfer functions, as has been done in temperate regions (e.g. Haghverdi et al., 2015 and Herbst and Diekkru, 2006). This would reduce labor costs – and time – and decrease the prediction error. Pachepsky et al. (1999) and (Botula et al., 2014) reviewed several approaches to predict soil water retention in mineral soils of temperate and tropical regions using pedotransfer functions. Among the most common predictors are bulk density, particle size distribution, organic carbon content, soil water potential, the geometry of the pore network, specific surface area, and cation exchange capacity. However, in the case of soils of Páramo ecosystems, new functions need to be developed that account for the role of soil organic matter and the particular mineralogy and chemistry that characterize them.

4. Conclusions

Through a precise methodological insight, we portrayed the large variability inherent in tuning and predictor selection using random forest. Due to the applied resampling methodology for model assessment via repeated 10-fold cross-validation an ensemble of 50 models is trained on different data subsets. As a consequence, any spatial prediction needs to be realized based on all 50 models. To guarantee for the robustness of the selected model parameters and predictors for each of the models, resampling was used for model tuning and predictor selection, respectively. Reporting the applied tuning, predictor selection, and resampling methods should become more widespread, because of their strong effect on the uncertainty of predictions and on the selection of important predictor variables.

Predictor selection helped in reducing model complexity aiding model interpretation. The most important predictors were altitude, positive openness – although just at pF 0.5 – and vegetation indices. On the regionalized maps, low median θ values were consistently observed for all pF values in the valley of the Quinuas River. This is likely due to a strong altitudinal control on major vegetation changes and land use, which are reflected in the vegetation indices. Due to the high variability in the small dataset, it is particularly these areas which display a high IQR due to the fact that mineral soils could not be separated from organic soils when adapting the random forest models.

The adapted models performed the best predicting θ between ca.

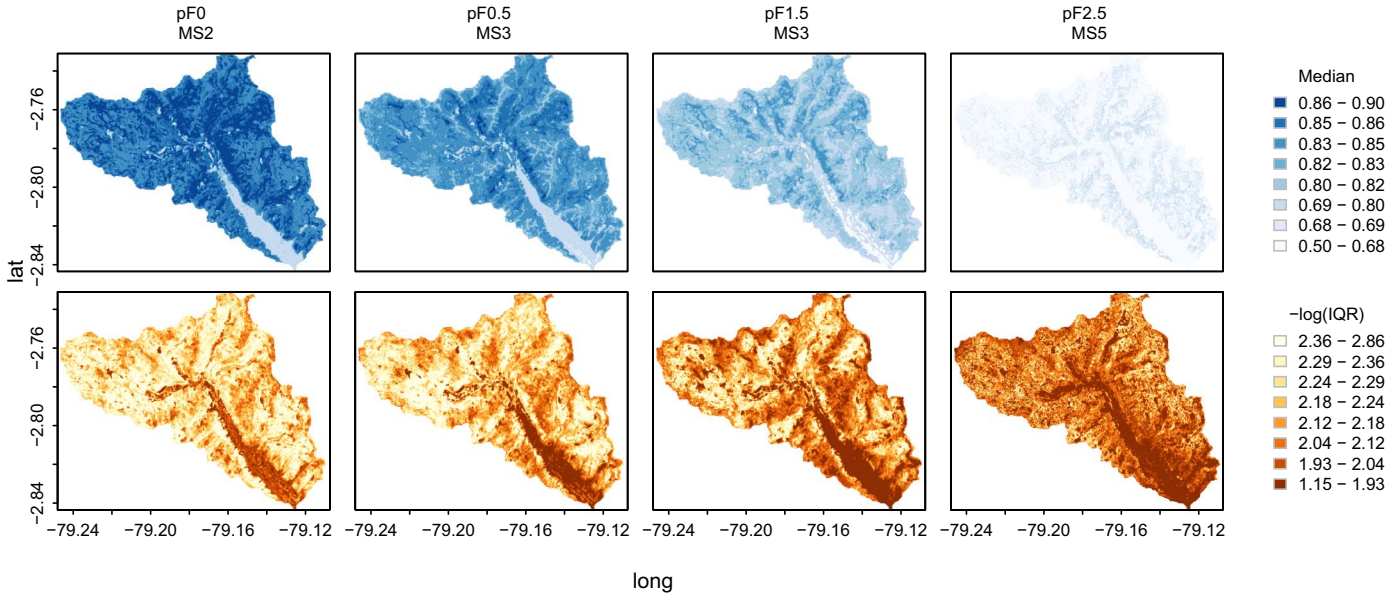

Fig. 11. Spatial predictions of the volumetric water contents (θ) at the different pF values based on the best model setup (MS). On the top row, each of the maps illustrates the median of the 50 predictions. On the bottom row, the corresponding log(IQR) is shown. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

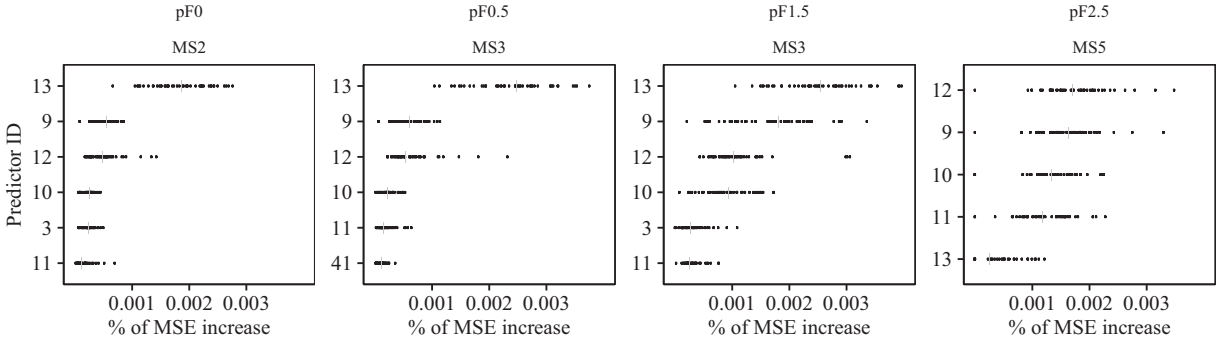

Fig. 12. Permutation importance scores of predictor variables resulting from the best model setup (MS) for each pF. Each dot represents the importance attributed to a predictor in a model run. The grey vertical lines represent the median importance of a predictor and were used to sort the predictors in descending order. Predictors with a median importance below 0.0001, i.e. below ca. 5% of the expected median of the most important predictor (around 0.002) are not shown.

0.55 and 0.9 (vol/vol). Within this range, the different model setups returned residuals whose relative magnitudes were ± 20%. The highest prediction errors at each pF were obtained for low θ and correlated with samples of high bulk density and low carbon content. Additionally, higher residual values were observed with increasing pF. This points to active mechanisms of water retention above pF 0.5, which are not represented by our predictor variables.

Possible explanations for the trend of the residuals are underrepresentation of samples typical for mineral soils – and in parallel, the susceptibility of random forest to extreme observations – and missing predictors that reflect the chemistry of the soil and/or are less time-dependent. Furthermore, data mining has shown that particularly for Páramos soils, there is no clear distinction between organic and mineral soils. The common differentiation due to the soil's organic carbon content subdivides the soil continuum by mere definition.

Although we regard our methodology as transferable to other Páramo regions, future work should strive to reduce the uncertainty of the predictions by improving any of the mentioned factors.

Acknowledgements

This research was funded by the German Research Foundation (DFG) as part of the Platform for Biodiversity and Ecosystem Monitoring and Research in South Ecuador (PAK 823, PAK 824 & PAK 825, LI 2360/1-1). Logistic support by the NGO Nature and Culture International (NCI) and the municipal public agency ETAPA is gratefully acknowledged.

References

Ambroise, C., McLachlan, G.J., 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. U. S. A. 99, 6562–6566. http://dx.doi.org/10.1073/pnas.102102699.

Arnalds, O., Buurman, P., Bartoli, F., Stoops, G., Garcia-ROdeja, E., 2007. Soils of Volcanic Regions in Europe.pdf. Springer.

Arroyo, M.T.K., Cavieres, L.A., Fuego, M., 2013. High-Elevation Andean Ecosystems. In: Encyclopedia of Biodiversity, Second Edition. pp. 96–110. http://dx.doi.org/10.1016/B978-0-12-384719-5.00428-7.

Asbjornsen, H., Manson, R., Scullion, J., Holwerda, F., Muñoz Villers, L., Alvarado, S., Geissert, D., Dawson, T., Bruijnzeel, L.A., McDonnell, J.J., 2017. Interactions between payments for hydrologic services, landowner decisions, and ecohydrological consequences: Synergies and disconnection in the 16 cloud forest zone of central Veracruz Mexico. Ecol. Soc. 22http://dx.doi.org/10.5751/ES-09144-220225. Art 25.

Barberi, F., Coltelli, M., Ferrara, G., Innocenti, F., Navarro, J.M., Santacroce, R., 1988. Plio-quaternary volcanism in Ecuador. Geol. Mag. 125, 1–14. http://dx.doi.org/10.1017/S0016756800009328.

Baret, F., Guyot, G., 1991. Potentials and limits of vegetation indexes for Lai and Apar assessment. Remote Sens. Environ. 35, 161–173.

Blume, H.-P., Brümer, G.W., Horn, R., Kandeler, E., Kögel-Knabner, I., Kretzschmar, R., Stahr, K., Wilke, B.-M., 2010. Lehrbuch der Bodenkunde. Soil Sci. http://dx.doi.org/10.1097/00010694-197703000-00015.

Boelter, D.H., 1966. Important physical properties of peat materials. In: Proceedings, Third Int. Peat Congr, pp. 150–154.
Boelter, D.H., 1969. Physical properties of peats as related to degree of decomposition. Soil Sci. Soc. Am. J. 33, 606. http://dx.doi.org/10.2136/sssaj1969.03615995003300040033x.
Böhner, J., Antonic, O., 2009. Land-surface parameters specific for topo-climatology. In: Hengl, T., Reuter, H.I. (Eds.), Geomorphometry: Concepts, Software, Applications. Elsevier, pp. 195–226.
Böhner, J., Selige, T., 2006. Spatial prediction of soil attributes using terrain analysis and climate regionalisation. In: SAGA - Anal. Model. Appl. Göttinger Aeographische Abhandlungen. 115. pp. 13–28.
Botula, Y.D., Van Ranst, E., Cornelis, W.M., 2014. Pedotransfer functions to predict water retention for soils of the humid tropics: a review. Rev. Bras. Ciênc. Solo 38, 679–698. http://dx.doi.org/10.1590/S0100-06832014000300001.
Breiman, L., 2001. Random forests. Mach. Learn. 5–32.
Breiman, L., 2003. Manual on Setting Up, Using, and Understanding Random Forests V4.0. (unpublished manuscript).
Buytaert, W., Wyseure, G., De Bièvre, B., Deckers, J., 2005. The effect of land-use changes on the hydrological behaviour of Histic andosols in south Ecuador. Hydrol. Process. 19, 3985–3997. http://dx.doi.org/10.1002/hyp.5867.
Buytaert, W., Célleri, R., De Bièvre, B., Cisneros, F., Wyseure, G., Deckers, J., Hofstede, R., 2006a. Human impact on the hydrology of the Andean paramos. Earth Sci. Rev. 79, 53–72. http://dx.doi.org/10.1016/j.earscirev.2006.06.002.
Buytaert, W., Celleri, R., Willems, P., De Bièvre, B., Wyseure, G., 2006b. Spatial and temporal rainfall variability in mountainous areas: a case study from the south Ecuadorian Andes. J. Hydrol. 329, 413–421. http://dx.doi.org/10.1016/j.jhydrol.2006.02.031.
Buytaert, W., Deckers, J., Wyseure, G., 2006c. Description and classification of non-allophanic andosols in south Ecuadorian alpine grasslands (páramo). Geomorphology 73, 207–221. http://dx.doi.org/10.1016/j.geomorph.2005.06.012.
Buytaert, W., Deckers, J., Wyseure, G., 2007. Regional variability of volcanic ash soils in south Ecuador: the relation with parent material, climate and land use. Catena 70, 143–154. http://dx.doi.org/10.1016/j.catena.2006.08.003.
Carlson, T.N., Ripley, D.A., 1997. On the relation between NDVI, fractional vegetation cover, and leaf area index. Remote Sens. Environ. 62, 241–252. http://dx.doi.org/10.1016/S0034-4257(97)00104-1.
Carrillo-Rojas, G., Silva, B., Córdova, M., Célleri, R., Bendix, J., 2016. Dynamic mapping of evapotranspiration using an energy balance-based model over an andean páramo catchment of southern ecuador. Remote Sens. 8. http://dx.doi.org/10.3390/rs8020160.
Casella, G., Fienberg, S., Olkin, I., 2006. An Introduction to Statistical Learning. Springerhttp://dx.doi.org/10.1016/j.peva.2007.06.006.
Celleri, R., Willems, P., Buytaert, W., Feyen, J., 2007. Space-time rainfall variability in the Paute basin, Ecuadorian Andes. Hydrol. Process. 21, 3316–3327. http://dx.doi.org/10.1002/hyp.6575.
Conrad, O., 2012. SAGA-GIS module library documentation (v2.2.3). In: Module Valley Depth.
Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner, J., 2015. System for automated geoscientific analyses (SAGA) v. 2.1.4. Geosci. Model Dev. 8, 1991–2007. http://dx.doi.org/10.5194/gmd-8-1991-2015.
Córdova, M., Célleri, R., Shellito, C.J., Orellana-Alvear, J., Abril, A., Carrillo-Rojas, G., 2016. Near-surface air temperature lapse rate over complex terrain in the Southern Ecuadorian Andes: implications for temperature mapping. Arct. Antarct. Alp. Res. 48, 678–684. http://dx.doi.org/10.1657/AAAR0015-077.
Crespo, P.J., Feyen, J., Buytaert, W., Bücker, A., Breuer, L., Frede, H.-G., Ramírez, M., 2011. Identifying controls of the rainfall – runoff response of small catchments in the tropical Andes (Ecuador). J. Hydrol. 407, 164–174. http://dx.doi.org/10.1016/j.jhydrol.2011.07.021.
Dangles, O., Rabatel, A., Kraemer, M., Zeballos, G., Soruco, A., Jacobsen, D., Anthelme, F., 2017. Ecosystem sentinels for climate change? Evidence of wetland cover changes over the last 30 years in the tropical Andes. PLoS One 12, e0175814. http://dx.doi.org/10.1371/journal.pone.0175814.
Díaz-Uriarte, R., Alvarez de Andrés, S., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinf. 7, 3. http://dx.doi.org/10.1186/1471-2105-7-3.
Durner, W., Lipsius, K., Durner, W., Lipsius, K., 2005. Determining soil hydraulic properties. In: Encyclopedia of Hydrological Sciences. John Wiley & Sons, Ltd, Chichester, UK. http://dx.doi.org/10.1002/0470848944.hsa077b.
FAO (Ed.), 2006. Guidelines for Soil Description, 4th ed. FAO, Rome.
Fox, G., Sabbagh, G., Searcy, S., Yang, C., 2004. An automated soil line identification routine for remotely sensed images. Soil Sci. Soc. Am. J. 68, 1326. http://dx.doi.org/10.2136/sssaj2004.1326.
Gao, B.C., 1996. NDWI - a normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens. Environ. 58, 257–266. http://dx.doi.org/10.1016/S0034-4257(96)00067-3.
Garfagnoli, F., Ciampalini, A., Moretti, S., Chiarantini, L., Vettori, S., 2013. Quantitative mapping of clay minerals using airborne imaging spectroscopy: new data on mugello (Italy) from SIM-GA prototypal sensor. Eur. J. Remote Sens. 46, 1–17. http://dx.doi.org/10.5721/EuJRS20134601.
Genuer, R., Poggi, J., Tuleau-malot, C., Genuer, R., Poggi, J., Variable, C.T., Genuer, R., Poggi, J., Tuleau-malot, C., 2010. Variable selection using random forests. Pattern Recogn. Lett. 31, 2225–2236.
Geroy, I.J., Gribb, M.M., Marshall, H.P., Chandler, D.G., Benner, S.G., Mcnamara, J.P., 2011. Aspect influences on soil water retention and storage. Hydrol. Process. 25, 3836–3842. http://dx.doi.org/10.1002/hyp.8281.
Gnatowski, T., Szatyłowicz, J., Brandyk, T., Kechavarzi, C., 2010. Hydraulic properties of fen peat soils in Poland. Geoderma 154, 188–195. http://dx.doi.org/10.1016/j.geoderma.2009.02.021.
Goslee, S.C., 2011. Analyzing remote sensing data in Rs: the landsat package. J. Stat. Softw. 43, 1–25. http://dx.doi.org/10.18637/jss.v043.i04.
Goslee, S., 2015. Landsat: radiometric and topographic correction of satellite imagery. In: R Package Version 1.0.8.
Goulard, M., Voltz, M., 1993. Geostatistical interpolation of curves: a case study in soil science. Geostat. Troia 92 (2), 805–816. http://dx.doi.org/10.1007/978-94-011-1739-5_64.
Gregorutti, B., Michel, B., Saint-Pierre, P., 2016. Correlation and variable importance in random forests. Stat. Comput. 1–20. http://dx.doi.org/10.1007/s11222-016-9646-1.
Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island - digital soil mapping using random forests analysis. Geoderma 146, 102–113. http://dx.doi.org/10.1016/j.geoderma.2008.05.008.
Grömping, U., 2009. Variable importance assessment in regression: linear regression versus random Forest. Am. Stat. 63, 308–319. http://dx.doi.org/10.1198/tast.2009.08199.
Gruber, S., Peckham, S., 2009. Land-surface parameters and objects in hydrology. Dev. Soil Sci. 33, 171–194. http://dx.doi.org/10.1016/S0166-2481(08)00007-X.
Guo, P.T., Li, M.F., Luo, W., Tang, Q.F., Liu, Z.W., Lin, Z.M., 2015. Digital mapping of soil organic matter for rubber plantation at regional scale: an application of random forest plus residuals kriging approach. Geoderma 237–238, 49–59. http://dx.doi.org/10.1016/j.geoderma.2014.08.009.
Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification. Mach. Learn. 46, 389–422. http://dx.doi.org/10.1108/03321640910919020.
Haghverdi, A., Leib, B.G., Washington-Allen, R.A., Ayers, P.D., Buschermohle, M.J., 2015. High-resolution prediction of soil available water content within the crop root zone. J. Hydrol. 530, 167–179. http://dx.doi.org/10.1016/j.jhydrol.2015.09.061.
Hastie, T., Tibshirani, R., Friedman, J., 2009a. The Elements of Statistical Learning, 2nd ed. Springer, New York. http://dx.doi.org/10.1007/978-0-387-98135-2.
Hastie, T., Tibshirani, R., Friedman, J., 2009b. The elements of statistical learning. In: Bayesian Forecasting and Dynamic Models, http://dx.doi.org/10.1007/b94608.
Hawkins, D.M., 2004. The Problem of Overfitting. pp. 1–12. http://dx.doi.org/10.1021/ci0342472.
Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., De Jesus, J.M., Tamene, L., Tondoh, J.E., 2015. Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions. PLoS One 10, 1–26. http://dx.doi.org/10.1371/journal.pone.0125814.
Hengl, T., Mendes de Jesus, J., Heuvelink, G.B.M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotic, A., et al., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS ONE 12 (2), e0169748. http://dx.doi.org/10.1371/journal.pone.0169748.
Herbst, M., Diekkru, B., 2006. Geostatistical co-regionalization of soil hydraulic properties in a micro-scale catchment using terrain attributes. Geoderma 132, 206–221. http://dx.doi.org/10.1016/j.geoderma.2005.05.008.
Hewelke, P., Gratowski, T., Hewelke, E., Żakowicz, S., Tyszka, J., 2015. Analysis of water retention capacity for selected forest soils in Poland. Pol. J. Environ. Stud. 24, 1013–1019. http://dx.doi.org/10.15244/pjoes/23259.
Hitziger, M., Ließ, M., 2014. Comparison of three supervised learning methods for digital soil mapping: application to a complex terrain in the Ecuadorian Andes. Appl. Environ. Soil Sci. 2014.
Hofstede, R., Segarra, P., Mena, P., 2003. Los páramos en el mundo: su diversidad y sus Habitantes, Los Páramos del Mundo.
Horta, A., Pereira, M.J., Ramos, T.B., 2014. Spatial modelling of soil hydraulic properties integrating different supports. J. Hydrol. 511, 1–9. http://dx.doi.org/10.1016/j.jhydrol.2014.01.027.
Iwahashi, J., Pike, R.J., 2007. Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. Geomorphology 86, 409–440. http://dx.doi.org/10.1016/j.geomorph.2006.09.012.
Jackson, T.J., Chen, D., Cosh, M., Li, F., Anderson, M., Walthall, C., Doriaswamy, P., Hunt, E.R., 2004. Vegetation water content mapping using Landsat data derived normalized difference water index for corn and soybeans. Remote Sens. Environ. 92, 475–482. http://dx.doi.org/10.1016/j.rse.2003.10.021.
Jenny, H., 1994. Factors of soil formation. In: Dover Edit. (Ed.), A System of Quantitative Pedology. Dover Publications, Inc, New York. http://dx.doi.org/10.1016/0016-7061(95)90014-4.
van der Keur, P., Iversen, B.V., 2006. Uncertainty in soil physical data at river basin scale - a review. Hydrol. Earth Syst. Sci. 10, 889–902. http://dx.doi.org/10.5194/hess-10-889-2006.
Khlosi, M., Cornelis, W.M., Douaik, A., van Genuchten, M.T., Gabriels, D., 2008. Performance evaluation of models that describe the soil water retention curve between saturation and oven dryness. Vadose Zo. J. 7, 87. http://dx.doi.org/10.2136/vzj2007.0099.
Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28, 1–26. http://dx.doi.org/10.1053/j.sodo.2009.03.002.
Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer, New York. http://dx.doi.org/10.1007/978-1-4614-6849-3.
Lee, S., Wolberg, G., Shin, S.Y., 1997. Scattered data interpolation with multilevel b-splines. IEEE Trans. Vis. Comput. Graph. 3, 228–244. http://dx.doi.org/10.1109/2945.620490.
Liaw, A., Wiener, M., 2002. Classification and regression by random forest. In: R News. 2. pp. 18–22. http://dx.doi.org/10.1177/154405910408300516.
Ließ, M., 2015. Sampling for regression-based digital soil mapping: closing the gap

between statistical desires and operational applicability. Spat. Stat. 13, 106–122. http://dx.doi.org/10.1016/j.spasta.2015.06.002.
Ließ, M., Hitziger, M., Huwe, B., 2014. The sloping mire soil-landscape of southern ecuador: influence of predictor resolution and model tuning on random forest predictions. Appl. Environ. Soil Sci. 2014. http://dx.doi.org/10.1155/2014/603132.
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. http://dx.doi.org/10.1371/journal.pone.0153673.
Maas, S.J., Rajan, N., 2010. Normalizing and converting image DC data using scatter plot matching. Remote Sens. 2, 1644–1661. http://dx.doi.org/10.3390/rs2071644.
McKenzie, N.J., Gessler, P.E., Ryan, P.J., O'connell, D.A., Wilson, J., Gallant, J., 2000. The role of terrain analysis in soil mapping. In: Terrain Analysis, pp. 245–265.
McKinney, B.A., Reif, D.M., Ritchie, M.D., Moore, J.H., 2006. Machine learning for detecting gene-gene interactions: a review. Appl. Bioinforma. 5, 77–88. http://dx.doi.org/10.2165/00822942-200605020-00002.
Mercado, L.M., Bellouin, N., Sitch, S., Boucher, O., Huntingford, C., Wild, M., Cox, P.M., 2009. Impact of changes in diffuse radiation on the global land carbon sink. Nature 458, 1014–1017. http://dx.doi.org/10.1038/nature07949.
Minaya, V., Corzo, G., Romero-Saltos, H., Van Der Kwast, J., Latinga, E., 2016. Altitudinal analysis of carbon stocks in the Antisana páramo, Ecuadorian Andes. J. Plant Ecol. 9, 553–563. http://dx.doi.org/10.1093/jpe/rtv073.
Möller, M., Volk, M., Friedrich, K., Lymburner, L., 2008. Placing soil-genesis and transport processes into a landscape context: a multiscale terrain-analysis approach. J. Plant Nutr. Soil Sci. 171, 419–430. http://dx.doi.org/10.1002/jpln.200625039.
Moore, I.D., Grayson, R.B., Ladson, A.R., 1991. Digital terrain modeling: a review of hydrological geomorphological and biological applications. Hydrol. Process. 5, 3–30. http://dx.doi.org/10.1002/hyp.3360050103.
Mosquera, G.M., Lazo, P.X., Célleri, R., Wilcox, B.P., Crespo, P., 2015. Runoff from tropical alpine grasslands increases with areal extent of wetlands. Catena 125, 120–128. http://dx.doi.org/10.1016/j.catena.2014.10.010.
Mulder, V.L., Lacoste, M., Richer-de-Forges, A.C., Martin, M.P., Arrouays, D., 2016. National versus global modelling the 3D distribution of soil organic carbon in mainland France. Geoderma 263, 16–34. http://dx.doi.org/10.1016/j.geoderma.2015.08.035.
Olaya, V., 2009. Basic land-surface parameters. In: Hengl, T., Reuter, H.I. (Eds.), Geomorphometry: Concepts, Software, Applications. Elsevier, pp. 141–169.
Oliveira, P.J.C., Davin, E.L., Seneviratne, S.I., 2010. Modeling the impacts of solar radiation partitioning into direct and diffuse fractions for the global water cycle. Geophys. Res. Abstr. 12, 4162.
Pachepsky, Y., Rawls, W.J., Timlin, D.J., 1999. The Current status of pedotransfer functions: their accuracy, reliability, and utility in field- and regional-scale modeling. In: Corwin, D.L., Loague, K., Ellsworth, T.R. (Eds.), Assessment of Non-Point Source Pollution in the Vadose Zone. Geophysical Monogr. 108. American Geophysical Union, Washington, DC, pp. 223–234. http://dx.doi.org/10.1029/GM108p0223.
Pachepsky, Y.A., Timlin, D.J., Rawls, W.J., 2001. Soil water retention as related to topographic variables. Soil Sci. Soc. Am. J. 65, 1787. http://dx.doi.org/10.2136/sssaj2001.1787.
Padrón, R.S., Wilcox, B.P., Crespo, P., Célleri, R., 2015. Rainfall in the Andean Páramo: new insights from high-resolution monitoring in southern Ecuador. J. Hydrometeorol. 16, 985–996. http://dx.doi.org/10.1175/JHM-D-14-0135.1.
Paloscia, S., Pettinato, S., Santi, E., Notarnicola, C., Pasolli, L., Reppucci, A., 2013. Soil moisture mapping using Sentinel-1 images: algorithm and preliminary validation. Remote Sens. Environ. 134, 234–248. http://dx.doi.org/10.1016/j.rse.2013.02.027.
Poulenard, J., Podwojewski, P., Herbillon, A.J., 2003. Characteristics of non-allophanic Andisols with hydric properties from the Ecuadorian páramos. Geoderma 117, 267–281. http://dx.doi.org/10.1016/S0016-7061(03)00128-9.
R Development Core Team, 2016. R: A Language and Environment for Statistical Computing. Version 3.3.1. 0 R Found. Stat. Comput, Vienna Austria ISBN 3–900051–07–0. https://doi.org/10.1038/sj.hdy.6800737.
Rezanezhad, F., Price, J.S., Quinton, W.L., Lennartz, B., Milojevic, T., Van Cappellen, P., 2016. Structure of peat soils and implications for water storage, flow and solute transport: a review update for geochemists. Chem. Geol. 429, 75–84. http://dx.doi.org/10.1016/j.chemgeo.2016.03.010.
Riley, S., De Gloria, S., Elliot, R., 1999. A terrain ruggedness that quantifies topographic heterogeneity. Intermt. J. Sci. 5, 23–27.
Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W., 1973. Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation. In: Prog. Rep. RSC 1978-1. 112 doi: 19740008955.
Saito, H., Seki, K., Šimůnek, J., 2008. Geostatistical modeling of spatial variability of water retention curves. Hydrol. Earth Syst. Sci. Discuss. 5, 2491–2522. http://dx.doi.org/10.5194/hessd-5-2491-2008.
Schneider, C., Flörke, M., De Stefano, L., Petersen-Perlman, J.D., 2016. Hydrological threats to riparian wetlands of international importance - a global quantitative and qualitative analysis. Hydrol. Earth Syst. Sci. Discuss. 1900, 1–35. http://dx.doi.org/10.5194/hess-2016-350.
Schwarzel, K., Renger, M., Sauerbrey, R., Wessolek, G., Schwärzel, K., Renger, M., Sauerbrey, R., Wessolek, G., 2002. Soil physical characteristics of peat soils. J. Plant Nutr. Soil Sci. Fur Pflanzenernahrung Und Bodenkd. 165, 479–486. http://dx.doi.org/10.1002/1522-2624(200208)165:4<479::aid-jpln479>3.0.co;2-8.
Schwärzel, K., Šimůnek, J., Stoffregen, H., Wessolek, G., van Genuchten, M.T., 2006. Estimation of the unsaturated hydraulic conductivity of peat soils. Vadose Zo. J. 5, 628. http://dx.doi.org/10.2136/vzj2005.0061.
Seibert, J., Stendahl, J., Sørensen, R., 2007. Topographical influences on soil properties in boreal forests. Geoderma 141, 139–148. http://dx.doi.org/10.1016/j.geoderma.2007.05.013.
Shangguan, W., Hengl, T., Mendes de Jesus, J., Yuan, H., Dai, Y., 2016. Mapping the global depth to bedrock for land surface modeling. J. Adv. Model. Earth Syst. 9, 1–24. http://dx.doi.org/10.1002/2016MS000686.Received.
Sinowski, W., Scheinost, A.C., Auerswald, K., 1997. Regionalisation of soil water retention curves in a highly variable soilscape, II. Comparison of regionalisation procedures using a pedotransfer function. Geoderma 78, 145–159.
Sklenář, P., Jørgensen, P.M., 1999. Distribution patterns of paramo plants in Ecuador. J. Biogeogr. 26, 681–691. http://dx.doi.org/10.1046/j.1365-2699.1999.00324.x.
Steinberg, A., Chabrillat, S., Stevens, A., Segl, K., Foerster, S., 2016. Prediction of common surface soil properties based on Vis-NIR airborne and simulated EnMAP imaging spectroscopy data: prediction accuracy and influence of spatial resolution. Remote Sens. 8, 613. http://dx.doi.org/10.3390/rs8070613.
Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinf. 8, 25. http://dx.doi.org/10.1186/1471-2105-8-25.
Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinf. 9, 307. http://dx.doi.org/10.1186/1471-2105-9-307.
Suárez, E., Arcos, E., Moreno, C., Encalada, A.C., 2013. Influence of vegetation types and ground cover on soil water infiltration capacity in a high-altitude páramo ecosystem. Avances 5, B14–B21.
Thompson, J.A., Roecker, S., Grunwald, S., Owens, P.R., 2012. Digital Soil. Interactions with and Applications for Hydropedology, Hydropedology, Mapping. http://dx.doi.org/10.1016/B978-0-12-386941-8.00021-6.
Too, V.K., Omuto, C.T., Biamah, E.K., Obiero, J.P., 2014. Review of soil water retention characteristic (SWRC) models between saturation and oven dryness. Open J. Mod. Hydrol. 4, 173–182. http://dx.doi.org/10.4236/ojmh.2014.44017.
Touw, W.G., Bayjanov, J.R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., Sacha van Hijum, A.F.T., 2013. Data mining in the life science swith random forest: a walk in the park or lost in the jungle? Brief. Bioinform. 14, 315–326. http://dx.doi.org/10.1093/bib/bbs034.
Tuller, M., Or, D., 2005. Water retention and characteristic curve. In: Encyclopedia of Soils in the Environment. Elsevier, pp. 278–289.
USGS, 2013a. Landsat 8. In: Fact Sheet, pp. 3–6. http://dx.doi.org/10.1002/0471743984.vse9497.
USGS, 2013b. Using the USGS Landsat 8 Product [WWW Document]. URL http://landsat.usgs.gov/Landsat8_Using_Product.php. (accessed 4.13.17).
Veerman, G.J., Stolte, J., 1997. Determination of the water retention characteristic using the hanging water column. In: Stolte, J. (Ed.), Manual for Soil Physical Measurements. Technical Document. 37. pp. 77 Wageningen.
Viviroli, D., Dürr, H.H., Messerli, B., Meybeck, M., Weingartner, R., 2007. Mountains of the world, water towers for humanity: typology, mapping, and global significance. Water Resour. Res. 43, 1–13. http://dx.doi.org/10.1029/2006WR005653.
Voltz, M., Goulard, M., 1994. Spatial interpolation of soil-moisture retention curves. Geoderma 62, 109–123.
Weiss, R., Alm, J., Laiho, R., Laine, J., 1998. Modeling moisture retention in peat soils. Soil Sci. Soc. Am. J. 62, 305–313. http://dx.doi.org/10.2136/sssaj1998.03615995006200020002x.
Wiegand, C.L., Richardson, A.J., Escobar, D.E., Gerbermann, A.H., 1991. Vegetation indices in crop assessment. Remote Sens. Environ. 119, 105–119.
Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic matter stocks using random forest modeling in a semi-arid steppe ecosystem. Plant Soil 340, 7–24. http://dx.doi.org/10.1007/s11104-010-0425-z.
Wösten, J.H.M., Pachepsky, Y.A., Rawls, W.J., 2001. Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. J. Hydrol. 251, 123–150.
Wright, M.N., Ziegler, A., König, I.R., 2016. Do little interactions get lost in dark random forests? BMC Bioinf. 17, 145. http://dx.doi.org/10.1186/s12859-016-0995-8.
Xiong, X., Grunwald, S., Myers, D.B., Kim, J., Harris, W.G., Comerford, N.B., 2014. Environmental Modelling & Software Holistic environmental soil-landscape modeling of soil organic carbon. Environ. Model. Softw. http://dx.doi.org/10.1016/j.envsoft.2014.03.004.
Xu, L., Saatchi, S.S., Yang, Y., Yu, Y., White, L., 2016. Performance of non - parametric algorithms for spatial mapping of tropical forest structure. Carbon Balance Manag. 18–20. http://dx.doi.org/10.1186/s13021-016-0062-9.
Yang, W.-H., Clifford, D., Minasny, B., 2015. Mapping soil water retention curves via spatial Bayesian hierarchical models. J. Hydrol. 524, 768–779. http://dx.doi.org/10.1016/j.jhydrol.2015.03.029.
Yokoyama, R., Shlrasawa, M., Pike, R.J., 2002. Visualizing topography by openness: a new application of image processing to digital elevation models. Photogramm. Eng. Remote. Sens. 68, 257–265.
Zuur, A.F., Ieno, E.N., Elphick, C.S., 2010. A protocol for data exploration to avoid common statistical problems. Methods Ecol. Evol. 1, 3–14. http://dx.doi.org/10.1111/j.2041-210X.2009.00001.x.