Norwegian University
of Life Sciences

Master's Thesis 2018 60 ECTS

Pedotransfer Functions for Predicting Hydraulic Properties of Non-Allophanic Andosols and Histosols in the Páramo of Southern Ecuador

Karoline Hildre Spilling
Environment and Natural Resources
Faculty of Environmental Sciences and Natural Resource Management

Preface / forord / antecedentes

Denne masteroppgava representerer mitt siste skoleprosjekt som student på linja *miljø og naturressurser* ved Universitetet i Ås. Temaet for oppgaven fikk jeg av forskningsgruppa iDRHICA i Cuenca i Ecuador, og jeg har hatt tett samarbeid med dem gjennom hele forskningsprosessen.

Jeg vil først og fremst takke mine veiledere Trond Børresen og Ellen Sandberg for gode tilbakemeldinger og råd. Internettsamfunnet fortjener en stor takk for hjelp når jeg har vært frustrert på stadige feilmeldinger i R Studio. Så vil jeg takke mamma og pappa for at dere er så varme og gode og for at dere lagde mat og vaska klærne mine da jeg var hjemme i måneden før innlevering. Takk til gode venner og storebror i Skoftestad bokollektiv for at døra deres var åpen da jeg trengte et sted å bo siste ukene på Ås. Takk til fine klassevenner for godt selskap i løpet av de siste par åra. Og takk til sangkoret Noe Ganske Annet, for en uforglemmelig studietid på Ås! Gledens beger er ei ennå tømt, og jeg ser fram til å danse sammen med dere igjen.

Gracias a Rolando Celleri y a Patricio Crespo en especial del *Departamento de Recursos Hídricos y Ciencias Ambientales* de la Universidad de Cuenca, por el tema, por el espacio en la Quinta y por la salida a Zhurucay. Gracias a Fausto por tu conversa en la ofi, siempre me hiciste sentir en casa. Y a Fránklin, muchas gracias por las horas que has pasado ayudándome con todas mis inquietudes acerca de la base de datos.

Al final, gracias a tí mi querido esposo loco, hermoso, precioso, adorado. Carlos, sin tu apoyo incondicional no hubiese podido acabar nunca mi proyecto de estudio. Y gracias al Shambo Alejo, por siempre alegrarme el día con tus locuras y tus churos de borrego. Ahora por fin me puedo concentrar en nuestro proyecto de familia, lo que vivimos y que viviremos en una futura chosita junto a min viejos, cholito lindo…

Ås, 14 mai 2018

Karoline Hildre Spilling

Summary

The *páramo* grasslands of southern Ecuador is a source of continuous clean water supply for downstream communities and ecosystems all the way to the Amazonian rainforests or the Pacific coast. Still, knowledge about environmental processes and interactions in the páramo is limited. The dominating soils covering the páramo of south Ecuador are extremely organic volcanic ash soils that are characterized by low density and a strong water retaining ability. A deep understanding of the unique hydraulic properties of these soils is necessary for reliable modelling of the páramo hydrology. However, measurement of soil hydraulic properties is time-consuming, costly and impractical for large-scale modelling, and simple estimation of the necessary variables using *pedotransfer functions* (PTFs) often gives a good enough approximation of field conditions. In this thesis, PTFs for predicting six water retention points, available water capacity and saturated hydraulic conductivity were developed for Andosols and Histosols in the páramo of southern Ecuador. In addition, a selection of existing PTFs were evaluated on the dataset of this thesis. Two statistical approaches were used for the PTF development, *ordinary least squares linear regression* (OLS) and *random forest* (RF). Possible predictor variables were *bulk density*, *organic matter*, *soil depth*, *slope*, *vegetation cover* and *soil type*. Predictive performances of the resulting PTFs were overall satisfactory, and both the OLS and RF approach achieved test RMSEs below 0.05 in the low soil-water suction range. Predictions in the high suction range were less accurate, but test RMSEs were still below 0.08. The RF models predicted a little more accurately than the corresponding OLS models because of the ability of the RF approach at capturing complex interactions between variables, but the difference was not considerable. The test errors of the models predicting saturated hydraulic conductivity were not especially accurate, but they might be good enough as an alternative to field measurements. After testing the reliability of the functions on new data sets, the PTFs can become useful tools for hydrological modelling that helps us to get a better understanding of environmental processes in the páramo.

Table of contents

1 Introduction

The *páramo* is a neo-tropical high altitude grassland ecosystem, that in Ecuador ranges from the continuous tree line at about 3500 m above mean sea level (MAMSL) to the permanent snow line at about 5000 MAMSL (Hofstede et al., 2003). They are ecologically vulnerable areas of high biological and socioeconomical importance, providing a continuous flow of clean water to ecosystems and communities downstream. Large metropolitan areas in the Andes without easy access to groundwater resources, depend completely on the páramos for their water supply. The high stability of water release from the páramo is due to the unique water retaining ability of the volcanic ash soils, *Andosols*.

Very little historical data exists on the hydrology of the páramo (Crespo et al., 2011), hence any research of the area is valuable to improve the understanding of the ecosystem, regardless of the possible services the páramo may provide to the downstream communities. Various studies have shown that the drying of volcanic ash soils for agricultural purposes cause an irreversible reduction of the soils' water retention capacity (Buytaert et al., 2002; Buytaert et al., 2004; Karube & Abe, 1998; Nanzyo et al., 1993; Poulenard et al., 2001; Shoji et al., 1996). When extensive areas of the páramo are affected in this way, downstream communities are more vulnerable to dry periods or flooding and erosion during intense rainfall. In addition to agriculture, mining activity and climate change are all factors threatening the biodiversity and ecosystem services of the south Ecuadorian páramo.

Knowledge about how water behaves in soil is crucial for understanding environmental processes and for modelling fluxes of water, contaminants and energy. Unfortunately, field measurement of soil water properties is time consuming and costly, and a simple estimation of these properties is often enough for practical uses or for large-scale hydrological modelling. Predictive models that estimate certain soil properties from other more easily accessible data, are commonly called *pedotransfer functions* (PTFs). There are many existing PTFs in the literature, but no generic function exists that applies to all the world's soils. Hodnett and Tomasella (2002) found that their PTFs developed on tropical soils did not predict well for Andosols, and they identified a need of developing specific models for this soil group. In a recent review article on the challenges and perspectives of soil PTFs, van Looy et al. (2017) state that there is a substantial knowledge gap for volcanic ash soils and peat soils.

The objectives of this thesis are as follows

- To contribute to a better understanding of the relationships between soil hydraulic properties and other soil properties and environmental factors in the páramo of southern Ecuador
- To develop a parallel set of pedotransfer functions using two different statistical approaches, for the prediction of six points on the water retention curve, available water capacity and saturated hydraulic conductivity of Andosols and Histosols
- To evaluate the predictive performance of the developed functions on an independent data set and compare the two approaches used in the PTF development
- To evaluate the predictive performance of a relevant existing PTFs on the same data set and compare with the performance of the developed PTFs
- To fit the van Genuchten water retention curve on the observed and predicted water retention values and evaluate the closeness of the prediction curves to the curves fitted on the observed data

2 Background

2.1 Andosols and Histosols in the páramo

Soils of the Ecuadorian páramo are greatly influenced by volcanic activity and deposits of volcanic ash, which is characteristic for the soil group *Andosol* (IUSS Working Group WRB, 2015). Shoji et al. (1996) give a thorough description of how soils with this unique parent material are developed, and their insight is summarized in the following paragraph; Volcanic ash is fine-textured, highly porous and permeable, which are all properties that speed up chemical weathering and elements are released at a higher rate than the formation of crystalline minerals. Hence, non-crystalline material accumulates in the soil. In humid climates, this material is typically allophane and imogolite or Al-humus complexes, depending on the soil pH and OM characteristics. Al-humus complexes form in soils with a high content of organic matter and a pH of around 5 or less, where organic acids are the dominant proton donor. Allophane and imogolite form in soils with pH ranging from 5 to 7 and with a low content of complexing organic compounds. Iron in Andosols are mostly found in ferrihydrite oxides rather than in less stable Fe-humus complexes.

Andosols typically accumulate organic matter, and Shoji et al. (1996) cite a number of studies (Brahim, 1987; Tate & Theng, 1980; Tokashiki & Wada, 1975; Wada, 1977) giving partial explanations to this phenomenon: Al-toxicity and P-deficiency for microorganisms; sorption of biodegrading enzymes to free aluminium and iron; and finally steric effects due to complexation and sorption that keep OM functional groups from reaching the microorganisms. The fact that the climate of the páramo in southern Ecuador is cold and wet, makes OM accumulation even more dominant. Organic matter is stabilized in organometallic complexation or by adsorbing to allophane, imogolite or iron oxides. Andosols generally have a low bulk density due of the formation of highly porous and stable aggregate structures that ensure that even the highly weathered soils have a considerable percentage of macro pores (Nanzyo et al., 1993). They retain phosphorus and water very strongly.

On water-saturated valleys floors, organic matter is typically accumulated and form *Histosols* (IUSS Working Group WRB, 2015), peat soils with hardly any mineral material in the surface layers. The density and hydraulic properties of Histosols depend greatly on the degree of decomposition.

2.2 Soil hydraulic properties

The main hydraulic properties of the soil are its *water retention capacity* and its *hydraulic conductivity*. Together, these two soil properties describe how water behaves in a soil, and knowledge of the two is crucial for the understanding of ecosystems and ecosystem services, for irrigation planning in agriculture and for addressing environmental impacts of climate change or human activity like construction, mining or agriculture.

2.2.1 Water retention capacity

A soil's water retention capacity (WRC) is the moisture content, θ, as a function of the matric potential, or *soil-water suction*, ψ. The soil-water suction is a measure of how firmly the soil matrix holds on to the water as gravity, evapotranspiration and plant uptake want to transport it elsewhere. Suction is defined as a negative pressure potential, and common units are *Pascal*, bar or water column (head). In this thesis, I use the unit *pF*, defined as the negative logarithm of the matric potential given in hPa or cm water column. Soil WRC is often represented as a *soil moisture characteristics curve*, where an increase in suction is associated with a decrease in water content. An example can be seen in figure 1.

When the suction is zero (pF 0, or 0 kPa in the figure) the soil is saturated with water. For the loam soil in the example, this means that approximately 53 % of the volume of the soil is water. The volumetric water content θ at saturation or pF 0 is equal to the porosity of the soil.

The water content at a matric potential of pF 2 - 2.5 (10 - 33 kPa) is traditionally called the *field capacity* of the soil, it is the moisture that remains in the soil a few days after a thorough wetting. At this point, gravity has drained out water that was stored in the largest pores, and the remaining water is retained in the soil by adsorptive and capillary forces acting between the water molecules and the soil particles. Water retention in the low-suction range (pF 0 - pF 2.5) depends mainly on the structure of the soil and the pore size distribution.

The water content at around pF 4.2 (1500 kPa) is called the *wilting point*, defined traditionally as the lowest amount of soil moisture needed for plants to survive. At this point, water retained by capillary forces in the space between particles is mostly gone, and only strongly bound adsorbed water remains, which is not accessible for plant uptake. Hence, water retention at high soil-water suctions is highly influenced by the soil's specific surface and sorption sites.

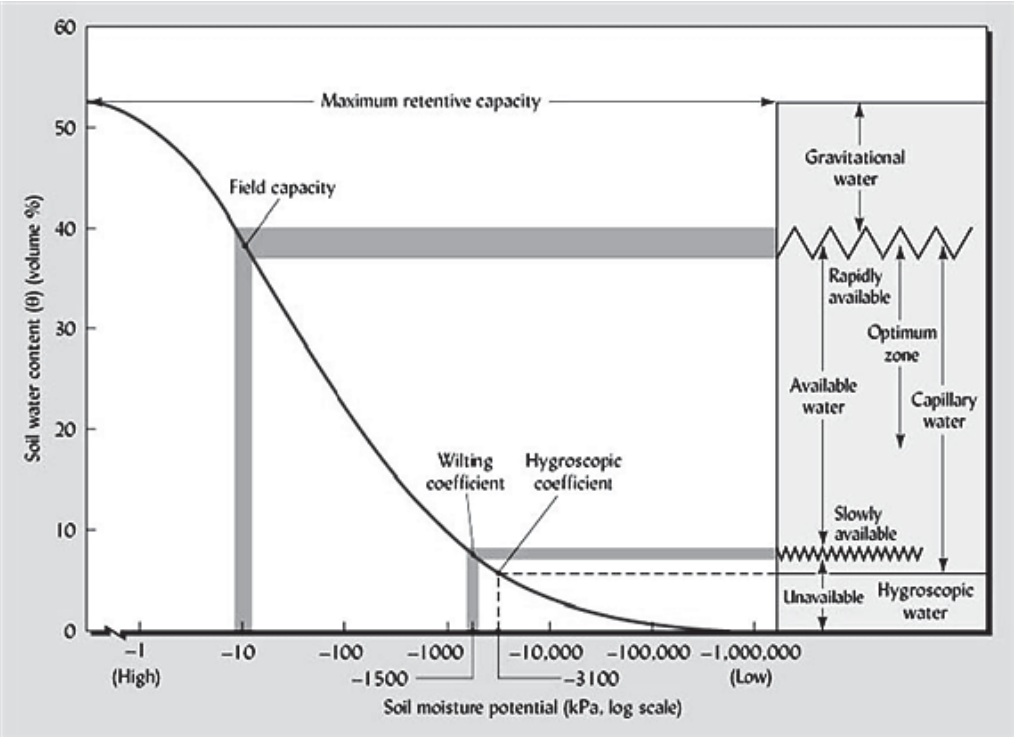

Figure 1. A typical soil moisture characteristics curve for a loam soil (figure 5.23 in Brady and Weil (2014))

Available water is the difference in water contents at field capacity and wilting point, and is defined as the amount of water that is available for uptake by plant roots. We say that water at matric potentials close to field capacity is rapidly available and the water at suctions close to the wilting point is slowly available. These terms or categorizations are faulty, and it is important to keep in mind that natural systems are more dynamic (Hillel, 2003). However, for agricultural purposes, the simplification is useful.

Some researchers have sought a universal analytical expression for describing the soil moisture characteristics curve, and well-known empirical expressions are the Brooks and Corey (1964), and the van Genuchten (1980) (VG). It is common practice to replace m parameter from the original publication with $m = 1 - 1/n$, to get fewer independent parameters, and thus we get the following formulation of the VG model,

$$\theta(\psi) = \theta_r + \frac{\theta_s - \theta_r}{[1 + (\alpha \cdot \psi)^n]^{1-1/n}}$$

where $\theta(\psi)$ is the water content at matric potential ψ, θ_r and θ_s are the residual and saturated water contents respectively, and α and n are scale- and curve-shaping parameters. The van Genuchten model has become the most used mathematical expression for describing water retention in soils, even though it may not be applicable to all soil types (Dettmann et al., 2014; Vereecken et al., 2010).

2.2.2 Saturated hydraulic conductivity

The saturated hydraulic conductivity, K_{sat}, is a soil property that describes how easily a fluid, in this case water, moves through the soil. Knowledge about the K_{sat} of a porous medium is essential for hydrological modelling, and in the calculation of flow and flow velocities of water and contaminants in saturated soils. It relates to the fluid's density, ρ, and its viscosity, μ, as well as the soil permeability, κ, in the following way,

$$K_{sat} = \frac{\kappa \cdot \rho \cdot g}{\mu}$$

where g is the acceleration of gravity, approximately $g = 9.81\ m/s^2$. The density and viscosity of water are often simplified to $\rho_w = 1\ kg/m^2$, and $\mu_w = 1 \cdot 10^{-3} Pa \cdot s$ respectively, even though both properties actually depend a little on temperature. This makes soil permeability the only variable factor in the formula, and it is the property that differentiates K_{sat} between porous media. Soil permeability depends on the pore space in the soil and on the interconnectivity of the pores, which in turn depends on the composition of the soil, on the way soil particles are organized and on the degree of compaction.

K_{sat} has the same SI unit as velocity, m/s, but it does not actually say anything about the flow velocity of water through the soil until it is considered together with the unitless factors *hydraulic head gradient* and the *effective porosity* (relative amount of interconnected pore space in the soil volume).

2.3 Pedotransfer functions

Water retention and hydraulic conductivity are both important soil properties that require a lot of time and resources to measure directly. This problem has since the beginning of the twentieth century made researchers look for quicker and cheaper ways of obtaining the needed information using more accessible information. There are plenty of examples of early studies estimating water retention from other soil properties (Briggs & McLane, 1907; Nielsen & Shaw, 1958; Riley, 1979; Salter & Williams, 1969; Veihmeyer & Hendrickson, 1927), and in the late 1980s the term *pedotransfer function* (PTF) was introduced describing a predictive model that *translates data we have into what we need* (Bouma & Lanen, 1987; Bouma, 1989). Soil properties like texture, geomorphology, horizon designations, organic matter content and bulk density all say something about the pore space and composition of the soil, and can possibly be linked to the way in which water behaves in the soil. Since the introduction of the PTFs, the interest in the idea of simply estimating soil moisture characteristics has increased with the increased use of computer modelling over large areas, where direct measurement becomes impractical.

Pedotransfer functions can be both continuous or categorical (Wösten et al., 1995), where the latter are based on groups with characteristics that influence hydraulic properties of the soil, like soil type, soil horizon, or textural class. Continuous PTFs give unique estimates based on continuous numerical data, like bulk density or percentages of clay, silt and sand in the soil.

2.3.1 Point vs. parametric PTFs and VG model applicability on studied soils

There are two approaches for the development of continuous PTFs for water retention, namely *point* and *parametric*. Point PTFs estimate water content at certain matric potentials, typically at saturation (pF 0), field capacity (pF ~ 2) and wilting point (pF ~ 4.2). The parametric PTFs predict the unknown parameters in an analytical equation to describe a continuous soil moisture characteristics curve. Very often the parameters $\theta_r, \theta_r, \alpha$ and n in the van Genuchten model are estimated. Dettmann et al. (2014) questions the applicability of classical analytical expressions such as the van Genuchten equation on Histosols and other soils with a high organic matter content, since these expressions are mainly developed on mineral soils that have completely different soil water dynamics. In addition, volcanic ash soils have very unique water retention properties, as confirmed by Hodnett and Tomasella (2002) who developed parametric PTFs to predict VG parameters on a variety of soils in the tropics. They found that the resulting PTFs were not reliable at all for Andosols, and recommended the development of specific parametric PTFs for these soils. This was done by Borja (2006) with acceptable results. Buytaert et al. (2005b) argues that a simple linear or semilogarithmic model might be better for describing the water retention curve in Andosols. Despite the difficulties, the VG expression has become a standard in most models describing flow in porous media (Vereecken et al., 2010), and an effort should be made to better understand how it relates to the special soils of this study.

Parametric PTFs are often preferred because of their ability to be used directly in modelling software that requires the parameters of an analytical equation. However, it is also possible to fit an analytical expression like the VG curve to water retention point estimates using software like SHYPFIT (Durner & Peters, 2009), RETC (Van Genuchten et al., 1991) or R (R Core Team, 2018). The parameters of the fitted curve can then be used as input in hydrological modelling software. Tomasella et al. (2003) compared the two approaches for obtaining a continuous soil moisture curve and found that using point PTFs and fitting a curve to the estimated points gave better results than directly estimating the analytical parameters. Water retention depends on different factors at different matric potentials, as mentioned in the previous section. This makes it difficult to estimate the curve parameters that apply for the curve in its totality, like n and α in the VG expression. Point PTFs involve the necessary predictors for explaining water retention at the different matric potentials.

2.3.2 Statistical methods for PTF development and evaluation

2.3.2.1 PTF development

PTF development was mainly ANOVA or linear regression analysis before modern tools for data mining became more common. Both methods requires a certain amount of *a priori* knowledge of the soil system and how properties are linked to obtain good predictive results. ANOVA-based class PTFs are normally based on soil texture class and/or soil type. They are simple, often in the form of look-up tables, and still widely applied, but they do not capture the dynamics inside each class. Moreover, the class predictions of soil water retention vary a lot between look-up tables, depending on the conditions of the soil used for the PTF development. Continuous PTFs based on linear regression are still simple and easy to interpret, but they capture more of the variation inside the categories if modelled well. A more thorough explanation of the regression method is given in a later paragraph.

Modern data mining techniques are becoming more common in the development of pedotransfer functions, and they require no previous knowledge to work well. Data mining methods are good at finding hidden structures in the data so all available information can be used in producing more accurate predictions. They are usually based on an input-output *black box* system, were information on soil basic properties is fed to the model as an *input*, and the model analyses the data and returns the predicted response, or the *output*. This approach makes the resulting models difficult to interpret compared to the more classical approaches. Data mining techniques that are commonly used for PTF development are *artificial neural networks*, *group method of data handling*, *support vector machines*, *k-nearest neighbour*-type algorithms, and *regression-/classification trees* and more sophisticated techniques based on regression-/classification trees, like *bagging*, *random forest* and *boosted random forest*. Breiman (2001), Pachepsky and Schaap (2004) and Nemes et al. (2006) explain the different PTF development methods in more detail, and the methods used in this thesis are described in the following paragraphs.

The ordinary least squares linear regression (OLS) is a classical approach for predicting a quantitative response from a given a set of predictor variables. It is frequently used for the development of pedotransfer functions, especially before the emerge of modern techniques, and the method gives simple, easily interpretable results. The linear regression model has the following form,

$$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon$$

where Y is the response variable, β_0 is the intercept of the model, $X_1, X_2, \dots, X_j$ are the predictor variables and $\beta_1, \beta_2, \dots, \beta_j$ are the coefficients of the predictor variables. The coefficient β_i of a

predictor can be interpreted as the change in the expected value of Y if the corresponding predictor shoul increase by 1 and all other predictors are held constant.

The final model term ε, represents the random prediction error associated with trying to simplify the real world. In linear regression, the idea is to minimize the random error to an acceptable level, and this is most commonly done by using the *ordinary least squares* approach: the difference between the observed and the estimated response is squared and the aim is to find predictor coefficients that minimize the sum of all these squared error terms, or the *residual sum of squares* (RSS).

Sometimes we want to include categorical data, for example gender or disease, as explanatory variables in our regression analysis. In these cases, we can use *dummy variables*, where numeric values (usually 0 and 1) are assigned to all the observations according to whether or not they belong in a certain group. If dummy variables are included in the final model, the resulting coefficients are multiplied by the assigned number of the dummy variable, which gives different intercepts and/or slopes to the linear model, depending on the group affiliation.

The **random forest** approach was presented by Leo Breiman in 2001 and has been used a lot for PTF development in recent years (Akpa et al., 2016; Koestel & Jorda, 2014; Sequeira et al., 2014). The method is based on *regression trees*, an approach that finds clusters in the data related to the response and makes a branched "tree" that ultimately leads to grouped predictions of the response variable. Figure 2 shows an example of a regression tree for the estimation of the water content at pF 2.4. Long branches indicate important splits, so the first split with bulk density (BD) lower or higher than 0.525 is by far the most significant in the example. The numbers at the end of the branches represent the estimates of water content at field capacity.

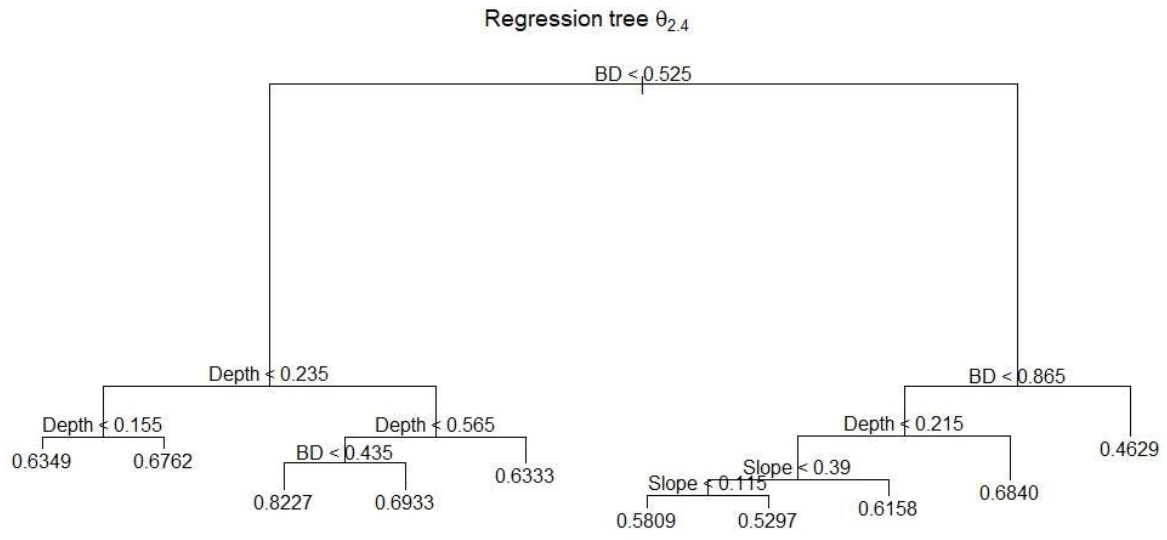

Figure 2. An example of a regression tree for the water content at pF 2.4

Regression trees are very intuitive and easy to interpret, and they can uncover non-linear structures and variable interactions in the data. However, they are not very accurate when used for prediction, compared to continuous approaches like regression (James et al., 2013).

Bagging and *random forest* are methods designed for predictive purposes. Both methods build hundreds of regression trees, each time with a different training set, using *bootstrapping*. The idea of bootstrapping is to use the already existing training data to make new, equally large data sets by randomly sampling with replacement from the original observations. Hence, some of the observations in the bootstrapped data sets will be repeated. Because of the bootstrapping approach, the many regression trees built in the bagging and random forest method will give different estimates of the response for a given set of predictor values. The trees in the model are heavily branched, which gives low bias, but the variance of the estimates of a single tree is high. By taking the average of the estimates from all the individual trees, the variance is reduced and this improves the predictive power of the model (James et al., 2013).

The difference between bagging and random forest is that the bagging method considers all possible predictors at each branch split. If one predictor is very dominating, the trees in the bagged model could end up looking very similar to one another, and the potential contribution of other, moderately strong predictors is not exploited. In a random forest model, there is a new set of m randomly chosen predictors considered for each branch split. This procedure makes the trees less correlated and the average response prediction of the trees less variable (James et al., 2013).

A random forest model is not as interpretable as a linear regression model, but it is possible to interpret to some degree by comparing the predictor variables' importance in the model. There are two common ways of calculating the importance of a variable: the mean *decrease in accuracy* and the mean *decrease in node impurity*. The decrease in accuracy of a variable is the decrease in mean squared error (MSE) of the predicted response if the variable in question was to take new random, but realistic values other than the original values. This can be simulated by shuffling the predictor variables randomly and running a random forest model on the "new" dataset. If a variable is important in the model, a change in its value will have a great effect on MSE of the model. If however a variable is unimportant, the effect on the response would be minor. The mean decrease in node impurity of a variable is a measure of how much the residual sum of squares (RSS) of each single tree in the model has decreased due to node splits over the variable. In other words, how much unexplained variance has been explained by the model after splitting over the variable in question. For both variable importance measures, larger values means that a variable is more important in the model.

2.3.2.2 PTF evaluation

There are many ways to evaluate a PTF's accuracy, i.e. the closeness of the model predictions to the real measurements. In most cases, there is an independent data set where the final models are tested. The prediction error on the test set can be expressed by the *correlation coefficient* (r) or the *coefficient of determination* (R^2), which are intuitive and well known statistics that give insight in how measurements and predictions are related and how much of the variance between them is explained by the pedotransfer model.

Another common statistic is the *mean error* (ME), which is the average prediction error, or the average difference between measured and predicted value. The ME is useful for detecting over- or underestimating of the response, as it will be either positive or negative. If the purpose is to find something like a standard deviance, it is common to use the *root mean squared error* (RMSE), given by the following formula,

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\varsigma' - \varsigma)^2}$$

where $\varsigma' - \varsigma$ is the difference between the predicted (ς') and the measured (ς) response and n is the number of observations. The statistic is easy to interpret as it has the same unit as the predicted response, unless the response is log transformed, in which case it will be unitless. When the test RMSE of a predictive model is given, it is easier for the end user to know how much variation to expect from the model predictions.

Sometimes we want to compare the performance of pedotransfer models on data originating from multiple sources with water retention measurements at different matric potentials. Or we want an idea of the validity of the model for predicting any point on the water retention curve. In these cases, it is useful to compare fitted WRC curves like the Van Genuchten curve discussed earlier. Tietje and Tapkenhinrichs (1993) defined a similar measure to the RMSE for comparing curves based on measured data and curves based on predictions; Instead of averaging a sum of squared prediction errors for given matric potentials, the *root mean squared difference* (RMSD) averages the integral of the squared errors over a defined range of matric potentials as follows,

$$RMSD = \sqrt{\frac{1}{b-a}\int_{a}^{b}(\varsigma' - \varsigma)^2 dh}$$

where a and b are the matric potentials at the limits of the defined integral, i.e. the range that is tested. Using the matric potentials at saturation (pF 0) and at the wilting point (pF 4.2) as the integral limits a and b, worked well for Tietje and Tapkenhinrichs, as it gave results on a similar

scale. However, the two measures RMSE and RMSD cannot be compared to each other directly, and there is no simple way of converting one into another for comparison (Vereecken et al., 2010).

After a PTF has been developed and validated using the same database, it is useful to address its *reliability*, i.e. evaluating the model's predictive performance on a completely different database with soils from other locations or projects.

2.3.3 Relevant existing PTFs

There are many PTFs available in the literature, and it can be practical to use existing functions if they prove to apply for the study area. In recent years, many publications have focused on the evaluation and comparison of existing models with varied results (Chen & Payne, 2001; Givi et al., 2004; Rawls et al., 2001). A widely used model is the Rosetta PTF software (G. Schaap et al., 2001), developed and validated on a large multinational database. Most of PTFs found in the literature, including Rosetta, are developed and tested in temperate climate regions. However, recent years have seen an increasing interest in the effects of land use change in the tropics on global climate, and a good understanding of the soil hydrology is key. Hodnett and Tomasella (2002) argues that the creation of a reliable universal PTF applicable to all soils worldwide, might not be possible due to the large regional variations in soil properties. They created a set of categorical and continuous PTFs for soils in tropical areas from the IGBP-DIS soil database, where 18 profiles were from Ecuador, and 27 profiles were Andosols. Their conclusion was that the texture class averages for bulk density and VG parameters in tropical regions were, in general, significantly different from the predictions that resulted when using class PTFs developed in temperate regions. The Andosols seemed to make things difficult for the tropical PTF development, with their exceptionally low bulk density, and the authors recommend separate PTFs for these soils. Batjes (1996) also found that Andosols and Histosols behaved differently in his generalisation of the hydraulic properties of the all the world's soils, and were studied separately.

Borja (2006) presented a set of PTFs on Andosols in Ecuador in his master's thesis at Cuenca University. The 87 data points he used for the model development originated from both the northern and the southern part of the Ecuadorian Andes, and the soil properties were quite different between the two regions. Soils from the north were more dense, they had coarser texture, less OM, lower WRC and higher K_{sat}. Borja used bulk density, soil texture and organic matter to develop point PTFs for six points on the soil moisture characteristics curve, parametric PTFs for the van Genuchten curve and saturated hydraulic conductivity. In some of the PTFs, he also included water contents at saturation (pF 0) or at high soil-water suctions (pF 3.48 and 4.18). He used both multiple linear regression and neural networks in the development of the models. Resulting PTFs were

validated on a test set of 13 observations, and test RMSE for the WRC point PTFs that did not include water contents as predictors, ranged from 0.094 to 0.122. When water contents were included, the lowest test RMSE dropped to 0.050. He found that using neural networks did not notably improve the predictive power of the model compared to classical linear regression, but he thinks that this can be explained by the relatively small number of observations used in the model development and -testing.

3 Material and methods

3.1 Study area

For the development of the functions in this thesis I used soil data originating from the páramo south-west of Cuenca city in the Azuay province in Ecuador at 2° 56' - 3° 06' S, and 79° 08' - 79° 18' W. The data was collected over a period of eight years (2008-2016), in four different projects at three locations in the Jubones and Paute river basins, namely *Zhurucay*, *Soldados* and *Tutupali* (figure 3). The altitude in the area ranges from 3400 to 4000 MAMSL, the mean annual temperature is 6 °C with an average relative humidity of about 90%. Annual rainfall is around 1300 mm, with 30% occurring as low intensity rainfall (drizzle); only 12% of the days are completely dry (Padrón et al., 2015).

Tussock grass (*Calamagrostis intermedia*) is the most common vegetation cover in the area, with moss and cushion bogs (*Plantago rigida*; *Xenophyllum humile*) in water saturated depressions. There are also occurring areas of the endemic woody Polylepis bushes (*Polylepis reticulata*) and pine tree plantations (*Pinus patula* and *radiata*). The photo in figure 4 shows the typical topography and natural vegetation in the study area. A pine tree plantation can be spotted on the hill on the left side.

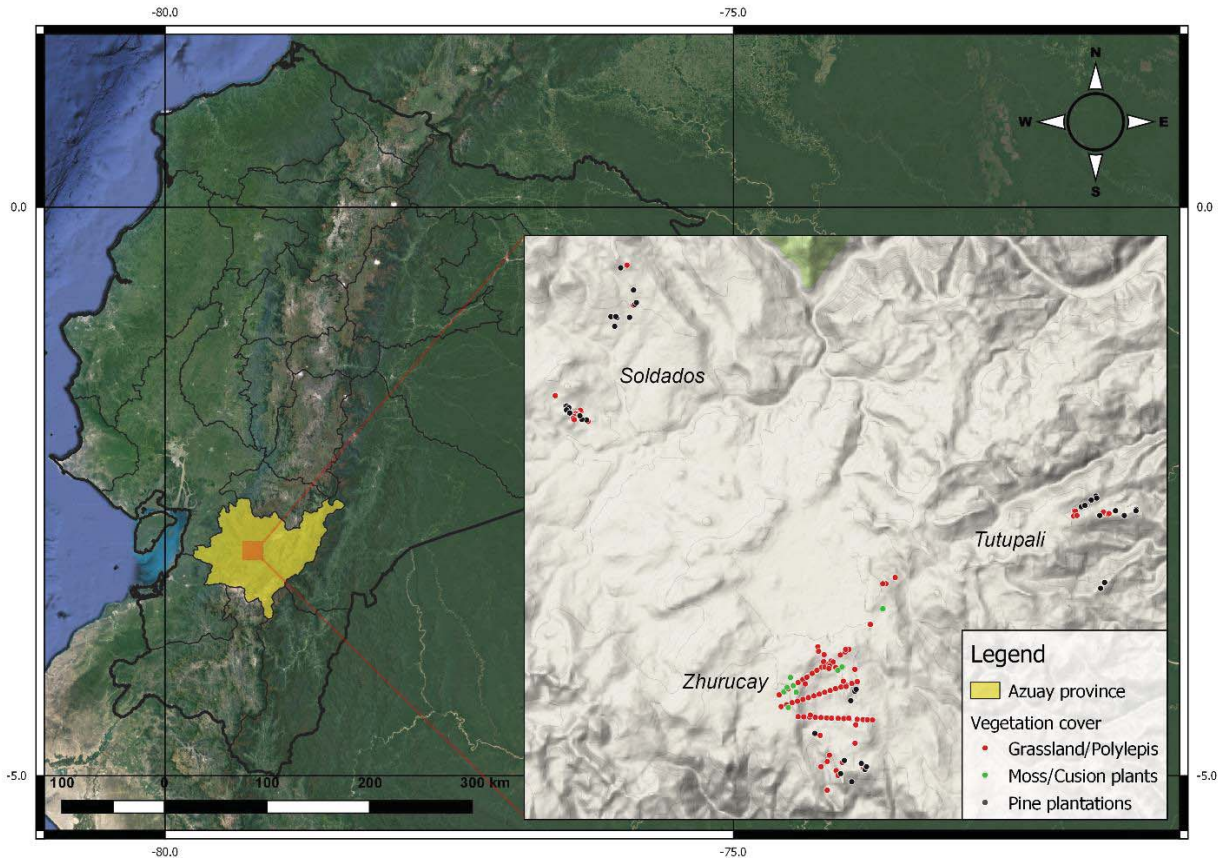

Figure 3. Study area with data point locations coloured after the respective vegetation cover on the site.

Figure 4. Private photo from the Zhurucay micro catchment.

The geology of the area is dominated by late Miocene to Plio-Pleistoscene acidic pyroclastics from the high volcanic activity during the rise of the Andes mountains (Buytaert et al., 2005a; Hungerbühler et al., 2002) and the topography is characterised by glacier formed valleys and plains. The area lies in the border zone of an area covered by recent ash deposits from Holocene volcanic activity in central Ecuador (Buytaert et al., 2004) and Andosol is the dominating soil type. Ash deposits are not very deep, due to the large distance to the recently active volcanos, and this coupled with the wet and cold environment has created a highly weathered soil rich in organic matter. The Andosols in the study site low in allophane and very high in organometallic complexes. Histosols are found on the water-saturated valley floors.

During the last decades, agricultural activity like controlled burnings, cattle grazing and establishments of pine tree plantations, has intensified in the study area (Quichimbo et al., 2012). The Canadian mining company INV Metals have the last years been doing preliminary feasibility studies in the area for gold extraction.

3.2 Projects and database

The 345 data points used to develop the functions in this thesis originates from four different projects which have been named *CT*, *AM*, *IM* and *TP*. Relevant objectives and conclusions from each of the projects will be presented shortly in this section. Table 1 presents maximum/minimum and centre values for the data by site and project.

3.2.1 CT – 14 data points from 10 soil profiles (3 Histosols, 7 Andosols)

Cajamarca and Tenorio (2008) did a descriptive study of the geomorphology and soils of the Quimsacocha páramo (Zhurucay area) for their undergraduate thesis at Cuenca University. They analysed water content at pF matric potentials 0, 1.57, 1.84, 2.04, 2.43, 2.73, 3.51 and 4.17. I have used their results from pFs 0, 1.57, 2.43, 3.51 and 4.17 for the development of the functions for the estimation of soil water contents at pFs 0, 1.5, 2.4, 3.4 and 4.2 respectively. Examples of dug soil profiles from the projects are given in figure 5.

Figure 5. Soil examples from the CT project in Zhurucay. Left: Histosol; Centre and right: Andosol. Photos taken from appendix 11 in the thesis of Cajamarca and Tenorio (2008).

3.2.2 AM – 61 data points from 43 soil profiles (9 Histosols, 34 Andosols)

Aucapiña and Marín (2014) studied how the landforms, categorized by the system in FAO's *Guidelines for soil description* (FAO, 2006; FAO, 2009), affect the hydraulic properties of the soil in Zhurucay. They analysed for water retention at pFs 0, 0.5, 1.5, 2.3, 3.4 and 4.2. Their results of the water content at pF 2.3 was used to develop the pedotransfer models for the water content at field capacity (pF 2.4) in this thesis.

The conclusions from their thesis was that the hydrophysical properties of the Zhurucay soil is strongly influenced by the landforms described in FAO (2006). They found a significant relationship between OM, BD and water retention at pFs 0, 0.5 and 1.5. In addition, they concluded that soils from valley bottoms generally have high water retention in the low soil-water suction range and drops to quite low water retention at high suction range. Soils from slopes or at the hill tops have a smaller difference between water content at low and high suctions. They did not find a significant relationship between the FAO landforms and the saturated hydraulic conductivity.

3.2.3 IC – 50 data points from 41 soil profiles (8 Histosols, 33 Andosols)

Irene Cárdenas (2014) did her undergraduate thesis work on the impact of pasture and pine plantations on the soil properties in both Zhurucay and Soldados. In her project there are only water retention data for pF matric potentials 0, 2.3 and 4.2. The hydraulic properties of the few data points with pastural activity proved in her thesis to be significantly different from the natural grassland and are thus not included in the development of the PTFs in this thesis.

In her thesis, Cárdenas found that the soils from pine tree plantations had a higher BD and K_{sat} than the natural grassland soils, and she believes this is due to the pine tree soils compacting, drying and cracking, as well as preferential subsurface flow along root systems. Cárdenas also found that the water content at field capacity was lower for pine tree soils than for the grassland soils.

3.2.4 TP – 220 data points from 110 soil profiles (110 Andosols)

The thesis work by Tapia and Pacheco (2015) was part of a still ongoing forest management project coordinated by *The University of Cuenca's Department of Water Resources and Environmental Sciences* (iDRHICA). They studied the effect of pine tree vegetation on properties of the andic surface horizon at two different depths. In addition to the data analysed in the mentioned thesis, some relevant available data points from the ongoing forest management project have been included in this thesis.

162 of the data points from this project are from six pine plantations established around 20 years ago; two in Soldados, three in Tutupali and one in Zhurucay. The soil around three representative trees in five 24 x 24 m blocks were studied for each plantation, except for one in Tutupali where data from only 2 blocks were included in this thesis because of altitude and vegetation differences. Originally, samples were taken at distances of 50 and 150 cm from the trees, but there were no statistically significant differences in relevant soil properties between the two distances (unpublished work from the ongoing project) and soil data for the two have been averaged for the PTF development in this thesis to avoid dependencies. The remaining 58 data points are control soils from the natural grasslands near the plantations. Soil samples were taken at two depths, 0-10 cm and 10-25 cm.

Tapia y Pacheco did not find significant differences in water retention between pine plantations and natural grasslands on an $\alpha = 0.05$ significance level, but the average K_{sat} proved to be significantly higher on the pine plantations, which they explain by preferential flow along the root systems.

Table 1. Summary of the total database (training and test set), by site and project

			Slope	Depth	BD	OM	θ_{pF0}	$\theta_{pF0.5}$	$\theta_{pF1.5}$	$\theta_{pF2.4}$	$\theta_{pF3.4}$	$\theta_{pF4.2}$	AWC	K_{sat}
			m m^{-1}	cm	g cm^{-3}	g g^{-1}	cm^3 cm^{-3}							cm h^{-1}
TOTAL		N	345	345	345	315	345	293	307	345	305	342	342	344
		Min - Max	0 – 0.67	5 - 98	0.06 - 1.35	0.01 – 0.94	0.47 - 0.96	0.46 - 0.92	0.44 - 0.91	0.35 - 0.95	0.16 - 0.72	0.08 - 0.63	0.00 - 0.80	0.03 - 16.9
		Median	0.20	20	0.53	0.30	0.75	0.73	0.71	0.61	0.46	0.41	0.20	1.20
		Mean	0.20	21	0.54	0.33	0.75	0.73	0.70	0.62	0.45	0.40	0.22	1.82
Zhurucay	**TP** (N = 42)	Min - Max	0.13 – 0.45	10 - 20	0.25 - 0.50	0.25 – 0.62	0.78 - 0.91	0.77 - 0.90	0.70 - 0.86	0.57 - 0.73	0.27 - 0.66	0.26 -0.58	0.09 - 0.42	0.12 - 2.89
		Median	0.22	15	0.37	0.40	0.84	0.82	0.77	0.65	0.45	0.41	0.22	0.86
		Mean	0.25	15	0.37	0.40	0.84	0.82	0.77	0.65	0.46	0.42	0.23	1.13
	AM (N = 61)	Min - Max	0.01 – 0.56	12 - 98	0.07 - 1.35	0.04 – 0.94	0.47 - 0.93	0.46 - 0.92	0.45 - 0.91	0.41 - 0.74	0.16 - 0.72	0.10 - 0.63	0.01 - 0.53	0.03 - 2.84
		Median	0.05	20	0.39	0.56	0.78	0.78	0.75	0.64	0.44	0.41	0.22	0.72
		Mean	0.11	33	0.45	0.50	0.77	0.76	0.73	0.63	0.45	0.39	0.24	0.87
	CT (N = 14)	Min - Max	0.02 – 0.38	5 - 86	0.06 - 1.11	0.01 – 0.80	0.63 - 0.91		0.59 - 0.91	0.57 - 0.90	0.16 - 0.65	0.08 - 0.53	0.26 - 0.80	0.38 - 4.10
		Median	0.06	40	0.32	0.35	0.87	-	0.86	0.84	0.52	0.38	0.40	1.13
		Mean	0.12	43	0.35	0.39	0.85		0.83	0.82	0.46	0.34	0.48	1.42
	IC (N = 38)	Min - Max	0 – 0.67	20 - 80	0.06 - 0.86	0.03 – 0.72	0.65 - 0.96			0.59 - 0.95		0.08 - 0.52	0.11 - 0.78	0.11 - 7.44
		Median	0.17	25	0.41	0.40	0.77	-	-	0.73	-	0.40	0.33	1.28
		Mean	0.20	32	0.42	0.37	0.78			0.75		0.36	0.39	1.41
Soldados	**TP** (N = 84)	Min - Max	0.10 – 0.36	10 - 20	0.46 - 1.04	0.10 – 0.37	0.52 - 0.79	0.50 - 0.76	0.44 - 0.75	0.35 - 0.65	0.20 - 0.57	0.14 - 0.51	0.00 - 0.45	0.27 - 5.04
		Median	0.18	15	0.67	0.23	0.69	0.68	0.63	0.51	0.44	0.40	0.11	1.48
		Mean	0.18	15	0.70	0.22	0.68	0.67	0.63	0.51	0.43	0.39	0.12	1.63
	IC (N = 12)	Min - Max	0.15 – 0.50	15 - 28	0.45 - 1.09	0.11 – 0.37	0.55 - 0.82	0.53 - 0.82	0.50 - 0.80	0.38 - 0.65	0.31 - 0.55	0.30 - 0.54	0.02 - 0.24	0.85 - 16.90
		Median	0.25	20	0.80	0.19	0.69	0.68	0.66	0.52	0.41	0.38	0.14	2.92
		Mean	0.27	20	0.75	0.20	0.70	0.69	0.65	0.53	0.42	0.39	0.14	5.56
Tutupali	**TP** (N = 94)	Min - Max	0.05 – 0.45	10 - 20	0.41 - 0.97	0.7 – 0.39	0.62 - 0.81	0.61 - 0.81	0.59 - 0.79	0.51 - 0.71	0.40 - 0.60	0.32 - 0.56	0.07 - 0.32	0.15 - 10.29
		Median	0.27	15	0.56	0.26	0.75	0.74	0.71	0.63	0.47	0.42	0.20	1.73
		Mean	0.27	15	0.59	0.25	0.74	0.73	0.70	0.61	0.47	0.42	0.20	2.66

3.3 Field and laboratory methods

Since the database origins from different projects, there is some variation in methodology, which is explained in the following sections of this chapter. More detailed information on procedures in each project can be found in the respective theses referenced in the previous section (all written in Spanish).

3.3.1 Soil sampling and profile description

Representative sites were chosen in all four projects for their respective objectives. Projects CT, AM and IC did a soil profile description for all data points, while project TP chose some representative sites for a full profile description. All profiles were described according to FAO's *Guidelines for soil description* (FAO, 2006; FAO, 2009). Undisturbed soil samples were collected using 100 cm^3 Kopecky cylinders (300 cm^3 cylinders in project CT) in the centre of all horizons for projects CT, AM and IC, and at depths 0-10 cm and 10-25 cm in project TP. Bags with disturbed soil were sampled to determine organic matter/carbon content and water contents at high soil-water suctions. In the TP project, only one disturbed soil sample was collected from each plantation block. Soil types were decided according to the characteristics described in FAO's second edition *World reference base for soil resources* (IUSS Working Group WRB, 2006).

3.3.2 Saturated hydraulic conductivity

All projects used the same methods to determine the saturated hydraulic conductivities at the studied soil depths and horizons. Three repetitions were done in the centre of all studied soil layers and the average values are used in this thesis.

In water-saturated horizons, the *auger-hole method* was used as described in detail by van Beers (1970) and Oosterbaan and Nijland (1994). The principle of the method is to excavate a hole down to the centre of the respective soil horizons using an auger. Then part of the water is removed and the rate at which the water rises is the hole is registered. The saturated hydraulic conductivity, K_{sat}, is calculated using the following equation,

$$K_{sat} = C\frac{\Delta H}{\Delta t}$$

where K_{sat} is the saturated hydraulic conductivity, Δt is the elapsed time between the first measurement of water level to the last measurement, ΔH is the change in water level during the time of measurement and C is a dimensionless geometry factor that depends on the depth of the hole below the water table and the distance from the hole bottom to a deeper, impermeable layer.

For unsaturated soil horizons, the *inversed auger-hole method* (also known as the *Porchet method*) was used as described by Oosterbaan and Nijland (1994). An auger is used to excavate a hole to

the depth of the centre of the soil horizon in question. Roots are cut, and the hole walls cleaned with a brush. Water is added in the hole until the surrounding soil is close to saturated, and the infiltration rate has stabilized. Then the hole is saturated with water up to the horizon limit and the rate at which the water sinks is registered. The saturated hydraulic conductivity is calculated using Darcy's law, $Q = v * A$, where the flow velocity v is set equal to K_{sat} and the area A is set equal to the area of the hole sidewalls and bottom, $A = 2\pi rh + \pi r^2$. Further transformation gives the following equation (see Oosterbaan and Nijland (1994) for details on the mathematical procedure,),

$$K_{sat} = 1.15r \frac{\log(h_0 + 0.5r) - \log(h_t + 0.5r)}{\Delta t}$$

where r is the radius of the auger-hole, h_0 is the water level in the hole at measurement start and h_t is the water level in the whole at time t.

3.3.3 Water retention capacity

In projects IC, AM and TP, the water content at saturation (pF 0) was obtained by leaving the undisturbed cylinder samples with a bandage and rubber band in a tray of water for approximately four weeks until they were completely saturated. The weights of the samples were registered.

Water retention at matric potentials pF 0.5 and pF 1.5 was analysed using the sandbox method in projects AM and TP. For this procedure, the saturated cylinder samples were transferred to a tray with water saturated sand. A suction of pF 0.5 (3.16 cm water column) was applied to the bottom of the sandbox for a week and the weight of the sample was registered. The cylinders were then returned to the sandbox and a suction of pF 1.5 (31.6 cm water column) was applied for another week before the sample was weighed again.

For the determination of water retention at field capacity, ceramic pressure plates were used in projects IC, AM and TP. The bandages and rubber bands were removed, and the cylinders placed on water saturated ceramic plates in a pressure chamber, making sure to obtain a good contact between the samples and the ceramic plates. Pressures of pF 2.52 (0.333 bar) in project TP, or pF 2.3 (0.2 bar) in projects AM and IC, were applied for a week and the weights were registered.

Finally, the cylinders were oven dried at 105 °C for 24 hours and the dry weigh of the soil samples were registered. The volumetric water contents at the different matric potentials were calculated dividing the weight differences between wet and dry samples, by the volume of the ring V_{ring} = 100 cm^3, assuming water has a density of 1 g cm^{-3} and considering the weight of the cylinders, the bandages and the rubber bands.

In project CT, the *multistep outflow* method was applied for the determination of water contents at suction higher than pF 3.4 (3 bar), as described by van Dam et al. (1994). In this method, an undisturbed cylinder soil sample (300 cm^3) is placed on a ceramic plate in a pressure cell and the sample is water saturated from below. Pressure is applied to the cell, while keeping the ceramic plate saturated, and this induces an unsaturated water flow through the soil. The water outflows and respective pressures are registered at given intervals as the pressure is increased. Finally, the samples were oven dried at 105 °C for 24 hours. The data was used as input to the program SHYP-FIT (Durner & Peters, 2009) that estimates the parameters of the van Genuchten equation and gives both water retention and hydraulic conductivity predictions.

All the projects used the same method for the determination of the water contents at matric potentials pF 3.4 and pF 4.2. A sub sample of the disturbed soil is sieved through a 2 mm sieve to remove roots and other coarse items; only the smallest particles are important for water retention in the high soil-water suction range. The purely organic soils were not sieved, but roots were removed. Water was mixed in with the sieved soil to make a shining paste, and the mix was covered and stored for 24 – 48 hours. Assigned rubber cylinders were placed on water saturated ceramic plates and they were filled with the soil mix and marked. The ceramic plates were placed in pressure chambers and suctions of pF 3.4 (3 bar) and pF 4.2 (15 bar) were applied in their respective chambers. After a week, the samples were taken out and weighed, before they were oven dried at 105 °C for 24 hours and weighed again. The gravimetric water content of each sample was calculated dividing the weight difference of the wet and dry sample, by the weight of the dry sample. The volumetric water content was then determined multiplying the gravimetric water content by the bulk density of the soil, assuming a water density of 1 g cm^{-3}.

3.3.4 Bulk density and organic matter content

To determine bulk density of the soil horizons, undisturbed cylinder samples were oven dried at 105 °C for 24 hours and then the dry samples were cooled and weighed. The dry bulk density is equal to the weight of the dry soil sample divided by the volume of the ring, V = 100 cm^3.

Organic matter was determined using the ignition method in the AM project. Disturbed sub samples from the mineral horizons were sieved through a 2 mm sieve and roots were removed for the organic horizons, but not sieved. The sub samples were oven dried on aluminium foil at 105 °C for 24 hours. 6 to 10 grams of the sub sample were put in a previously weighed and coded crucible and the weights of the crucible and the soil were registered. The crucibles with the soil was ignited at 430 °C for four hours before they were taken out to cool and weighed again (crucible and soil).

The organic matter fraction of the soil was determined by dividing the weight loss on ignition by the weight of the dry soil sample.

The rest of the data points have information on the carbon content, which in the acid soils of the study area is almost exclusively organic carbon. A classic factor of 1.724 has been applied in their respective analyses to convert the organic carbon content to organic matter content. However, Pribyl (2010) argues that this factor is too low in most cases, and that a factor of 2 is more accurate for almost all of the 24 studies of soils from all over the world. In the TP project, organic matter content was measured with the ignition method in the representative soil profiles. This data was compared to the carbon content measured from the same blocks and depths and the conversion factor turned out to be 2.098 (R^2 = 69 %). Since data from multiple projects are used in the PTF development, and the AM project used the ignition method, a conversion factor of 2 has been applied to the carbon data in the other projects to get a more accurate estimate of the organic matter contents.

3.4 Data analysis and PTF development

Data points originated from three different projects at three locations, which could lead to unwanted noise in the models, because of possible systematic differences in methods between the projects or geographical variation. Including *random effects* in a model can in many cases account for this kind of variance. This was tried in all models, but the random effects turned out to be insignificant and they were left out for simplicity.

Observations deeper than 100 cm or with soil types other than Andosols and Histosols were deleted due to scarcity of data. The hydraulic conductivity was log-transformed to achieve a more normal distribution of the observations. The natural logarithm was used, and the variable is denoted *logKsat*. The data set was divided in two parts; a training set with 75 % of the data for model development, and a test set with the remaining observations for model validation. The training set was again divided into ten subsets for cross validation. When using data from multiple sources to build predictive functions, Jorda et al. (2015) recommend using the different sources as subsets for cross validation to ensure an unbiased and more realistic model building. However, the distribution of the data points between the projects in this thesis was not even enough to perform a source-wise cross validation. Instead, the training-/test sets and the ten cross-validation subsets was balanced with respect to project and site to ensure a fair generalization of the data.

3.4.1 Choosing appropriate predictors

Soil texture is the most commonly used predictor in pedotransfer models, but it was not included in the PTFs of this thesis. While information about texture might be readily available from soil

databases in many countries or regions, this is not the case for the páramos of Ecuador. A particle size distribution analysis is time consuming, and as the purpose of pedotransfer models is to save time and resources, it does not make much sense to include texture in the models for this study area. Moreover, texture analysis of Andosols has proven to be problematic and texture data must be treated with caution, as the volcanic ash in combination with organic matter creates extra stable aggregates that are not easily dispersed (Buurman et al., 1997; Mizota & Van Reeuwijk, 1989). According to Nanzyo et al. (1993), the right way of dispersing non-crystalline clays after removing organic matter with hot H_2O_2 is to pass the sample through an ultrasonic treatment, adjust the pH and wash repeatedly with deionized water. In addition to the mentioned difficulties, the data on soil texture was far from complete in the database used in this thesis.

Information on *bulk density*, *vegetation cover*, *soil type*, *slope* and *depth* was complete in the data set, and these properties are linked to the pore size distribution, structure and composition of the soil, deciding its hydraulic functions. They are easy to measure or readily available and relevant for hydrological modelling on a larger scale using geographical information systems. These five mentioned properties were thus the chosen variables for the development of the pedotransfer models in this thesis. Vegetation cover was grouped in the following way, based on intuition and a quick look at the data:

1. Grassland or Polylepis (dominating natural vegetation cover in study area)
2. Moss or cushion plants (found in water saturated areas)
3. Pine tree plantations

The data for the *organic matter content* (OM) was incomplete, but the property was still used as a predictor in some models were it significantly improved the predictive power.

3.4.2 PTF development and evaluation

All data analysis was done in R, version 3.4.4 (R Core Team, 2018). Two modelling approaches were selected for the development of the PTFs in this thesis: *ordinary least squares linear regression* and *random forest*.

In the ordinary least squares model development, the categorical variables *vegetation* and *soil type* were converted to dummy variables. The R package `glmulti` was used to fit all possible linear models using the chosen predictors and their two-way interactions. A best model subset was produced based on the models' *Bayesian information criterion* (BIC), given by the following formula,

$$BIC = \frac{1}{n}(RSS + \log(n)\, d\hat{\sigma}^2)$$

where n is the number of observations, the RSS is the residual sum of squares, or random error, d is the number of predictors and $\hat{\sigma}^2$ is an estimate of the variance of measurement error of the response. The BIC is closely related to the *Akaike information criterion* (AIC) and *Mallow's Cp*, and is a good measure for finding a model that explains the data well without it being too complex, the BIC increases as the number of predictors d increases. A ten-fold cross validation was performed on the 50 models with the lowest BIC, and the average test RMSE was computed for all the models. Of the models with the lowest cross validation RMSE, the simplest one was chosen, it was fitted again to the whole training set, and the model assumptions were checked graphically. Finally, the model was used to predict the response on the separated test set and the final test RMSE was calculated.

In the **random forest** model development, it was not necessary to use dummy variables when including the categorical variables, as the method is based on regression trees, which is a grouping approach. The R package `randomForest` was used for the development of the RF models. A ten-fold cross validation was performed with each possible value of m predictors, to find the value that best exploited the information in the data without correlating the trees. The m value that gave the lowest cross validation RMSE after fitting 500 trees, was chosen for the final RF model, that was fitted on the whole training set. The m predictors are chosen randomly at each branch in every tree, so the R function `set.seed()` was used to ensure that the same results could be reproduced when running the script multiple times. The model was finally used to predict the response from the variable in the test set and the final test RMSE was calculated.

The predictive performance of the two approaches was evaluated and compared by the RMSE of the final model predictions on the test set. In the cases where dummy variables were included in an OLS model, total test RMSE was accompanied by separate RMSEs calculated using only the data in the test set belonging to the dummy variable category.

3.4.3 van Genuchten WRC curves

I chose not to develop parametric pedotransfer models that directly estimate the parameters θ_r, θ_r, n and α in the Van Genuchten equation, because of the difficulties mentioned in section 2.3.1. Instead, the R package `soilphysics` was used with point values for water retention to fit continuous curves and retrieve the associated VG parameters. If curves had problems converging, it was possible to manually adjust the initial parameter values which helped in some cases. Three curves were fitted for each set of observations in the test set, one from the measured water retention data, one from the RF estimates and one from the OLS estimates. The curve-fitting performance of the two methods was evaluated by comparing their average RMSD.

3.5 Evaluation of existing pedotransfer functions on test set

After an extensive search in the published literature, I did not manage to find any PTFs predicting soil hydraulic properties that did not involve soil texture, nor any other variable that was unavailable in the database used in this thesis. Most of the Ecuadorian PTFs developed by Borja (2006) included soil texture, but there was one that only included bulk density and some that included other water retention points. A relative selection of these are presented in table 2. Borja Ramón also developed PTFs using the *artificial neural network* (ANN) data mining approach with the same variables, but these functions were not available for evaluation. The published test RMSE of both approaches is presented in the table, to get an idea of the ANN models' performance compared to the linear regression models.

We see that the ANN PTFs did not perform better that the linear regression for water contents at pF1.98 an pF4.18, but they did better for both pF3.48 models. Borja Ramón developed his functions on a limited data set, using only 87 observations for development and 13 for validation. And observations were from both northern and southern Ecuador, were the soil properties were different. This may explain the relatively high test RMSEs for both methods in his publications.

The PTFs from table 2 were evaluated on the test set in this thesis, and their new test RMSEs were compared with RMSEs obtained from the developed functions. I have chosen not to include water retention points as predictor variables, but it is still interesting to see how well already existing PTFs perform on a new data set.

Table 1. Relevant PTFs developed by Borja Ramón on Ecuadorian Andosols and their performances in the publication

Code	Pedotransfer function	Test RMSE	Test RMSE ANN equivalent
MRLM2a	$\theta_{pF0} = 0.97984 - 0.38024\,BD$	0.1054	-
MRLM2b	$\theta_{pF1.98} = -0.15691 - 0.00214\,OM + 1.12902\,\theta_{pF0}$	0.0499	0.0592 (MRNA2b)
MRLM4e	$\theta_{pF3.48} = 0.06799 + 0.98129\,\theta_{pF4.18}$	0.0608	0.0461 (MRNA4e)
MRLM4f	$\theta_{pF4.18} = 0.00990 - 0.03696\,BD + 0.862956\,\theta_{pF3.48}$	0.0588	0.0681 (MRNA4f)

4 Results and discussion

4.1 Preliminary analysis of data in training set

Figure 6 presents the linear relationships between the variables in the training set (N=260), with Pearson correlation coefficients on the upper panel and scatterplots on the lower. Data points are coloured by the sampling site.

Bulk density is the predictor with the strongest correlation with water contents in the low suction range. The negative relationship changes as suction increases, until a positive correlation is established for very low BD. We see the same pattern with water retention and organic matter, where a positive correlation exists for OM contents below 50%. However, for higher OM contents, the curve flattens, and it even drops for high soil-water suctions. The available water capacity correlates with BD and OM and seems to increase exponentially for soils with very low BD.

From the figure, *slope* does not seem to have very strong linear relationships with any of the soil hydraulic variables, but are best correlated with water contents at high soil-water suctions, AWC and logKsat. Soil depth has a negative linear relationship with water retention at high soil-water suction, while AWC increases with depth. The saturated hydraulic conductivity is the most correlated to soil depth, with a clear negative relationship.

The boxplots in figure 7 show the trends of the soil hydraulic properties in the training set for every level of the categorical variables *vegetation cover* and *soil type*. We see that Histosol is the soil type

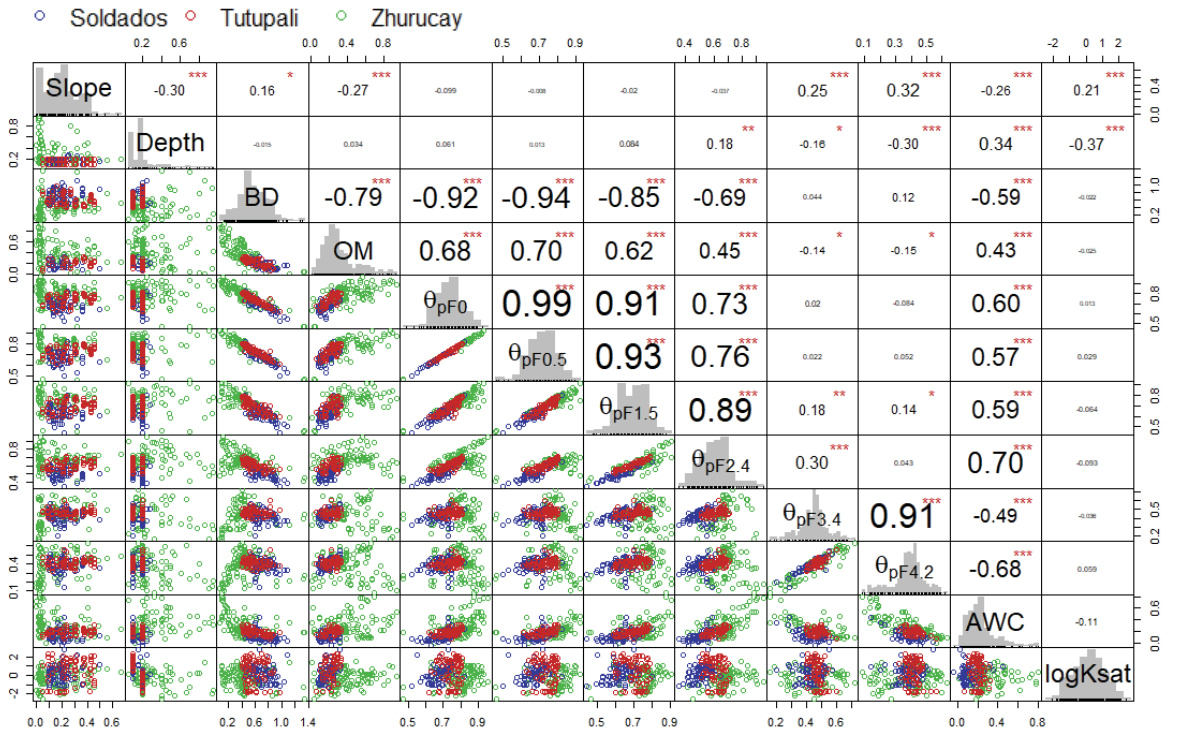

Figure 6. Training set scatterplots coloured by site and Pearson correlation coefficients between the chosen numeric variables. Grey histograms on diagonal show the frequency distribution of observations. α-significance levels: * 0.05 ** 0.01 *** 0.001

with the highest average water retention in the low suction range, while the water content drops below that of Andosols in the high soil-water suction range. The Andosol observations are many and quite scattered, but the average trend is that the difference between water content at saturation (pF 0) and water content at the wilting point (pF 4.2), is small compared to the Histosols, or to the loam soil example in figure 1. Hence, Andosols have a lower AWC, reflected in the boxplot figure *b*. It is hard to say something general about how the trends in K_{sat} differ between soil types, other than the observations are more scattered for Andosols than for Histosols.

Many of the observations with moss and cushion plant vegetation coincide with the Histosol observations, thus the two groups show similar tendencies in both water retention and hydraulic conductivity. Grassland/Polylepis observations are scattered, but the patterns are comparable to those of Andosols. The same can be said of the pine tree data, although it is much less scattered, at least for water retention. The average water content at field capacity (pF 2.4) is lowest in soils from pine tree plantations. Saturated hydraulic conductivity varies more for these soils than the soils belonging to the two other vegetation classes.

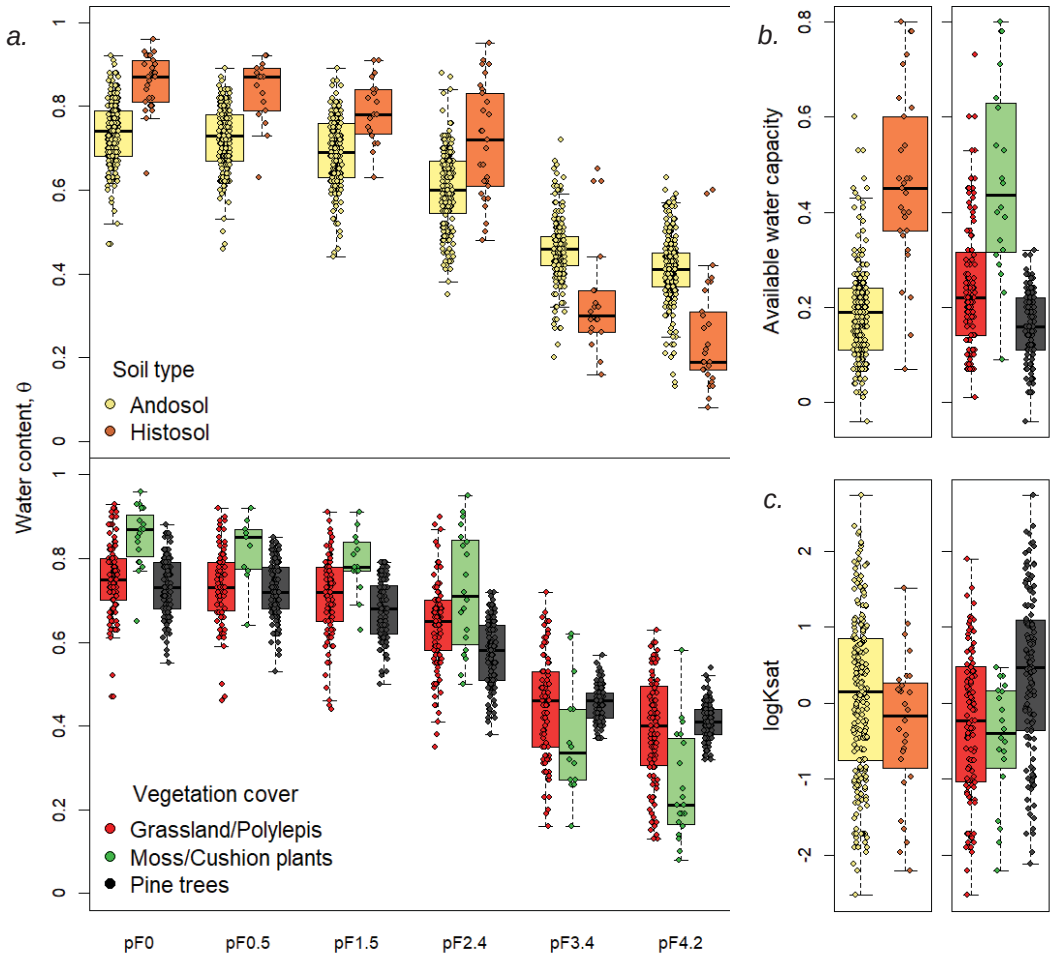

Figure 7. Boxplots showing trends on a: water retention, b: available water capacity and c: logKsat depending on *soil type* (yellow and orange plots) or the *vegetation* cover (red, green and black plots). The dots represent the training set observations.

4.2 Resulting pedotransfer functions and variable importance

4.2.1 OLS linear regression functions

The resulting OLS are presented in table 3. For some of the matric potentials, the models performed best when using dummy variables to create separate functions for Histosols or soils from pine plantations. The functions are coded with a parenthesis with a capital letter at the end: (H) for *Histosols*, (P) for *pine* tree soils and (R) for the *rest* of the cases. The dummy variable functions are part of the same linear model, but they are written out as separate functions for the sake of interpretability.

Organic matter content was an important predictor for water content at high suctions and for the AWC. OM was however not important enough to include in the models in the lower suction range. The effect of an increase of BD alone is negative in all cases except for K_{sat} on the pine plantations. Soil depth was a significant predictor for logKsat and for water retention at pF1.5, field capacity (pF 2.4) and for AWC. For water retention at both pF1.5, pF2.4 and AWC, *Histosols* proved to be a significant dummy variable, and are separated in a PTF apart. The *Pine* dummy variable was significant in all models except the low suction water retention models.

Table 2. OLS pedotransfer functions for estimating points on the water retention curve, available water capacity and the log of saturated hydraulic conductivity. H: Histosol; P: soil from pine plantation; R: not Histosol nor soil from pine plantation

Code	Pedotransfer function
OLS0	$\theta_{pF0} = 0.93849 - 0.34568\ BD$
OLS0.5	$\theta_{pF0.5} = 0.93265 - 0.35262\ BD$
OLS1.5(H)	$\theta_{pF1.5} = 0.81660 - 0.37152\ BD + 0.13370\ Depth + 0.08460\ Slope$
OLS1.5(R)	$\theta_{pF1.5} = 0.86537 - 0.37152\ BD + 0.13370\ Depth + 0.08460\ Slope$
OLS2.4(H)	$\theta_{pF2.4} = 0.78319 - 0.08317\ BD - 0.12070\ Depth$
OLS2.4(P)	$\theta_{pF2.4} = 0.75665 - 0.37555\ BD + 0.26297\ Depth$
OLS2.4(R)	$\theta_{pF2.4} = 0.78766 - 0.37555\ BD + 0.26297\ Depth$
OLS3.4(P)	$\theta_{pF3.4} = 0.61299 - 0.25920\ BD - 0.14891\ Slope - 0.75548\ OM + 1.25525\ BD \cdot OM + 0.75056\ Slope \cdot OM$
OLS3.4(R)	$\theta_{pF3.4} = 0.45504 - 0.12088\ BD - 0.14890\ Slope - 0.36857\ OM + 1.25525\ BD \cdot OM + 0.75056\ Slope \cdot OM$
OLS4.2(H)	$\theta_{pF4.2} = 0.26177 - 0.12330\ BD - 0.22864\ OM + 1.44139\ BD \cdot OM$
OLS4.2(P)	$\theta_{pF4.2} = 0.55066 - 0.30454\ BD - 0.62947\ OM + 144139\ BD \cdot OM$
OLS4.2(R)	$\theta_{pF4.2} = 0.34754 - 0.12330\ BD - 0.22864\ OM + 1.44139\ BD \cdot OM$
OLSAW(H)	$AWC = 0.76580 - 0.39703\ BD - 0.10580\ Depth - 0.12835\ OM - 0.85848\ BD \cdot OM$
OLSAW(P)	$AWC = 0.21722 - 0.13322\ BD + 0.22571\ Depth + 0.42841\ OM - 0.85848\ BD \cdot OM$
OLSAW(R)	$AWC = 0.56363 - 0.39703\ BD + 0.22571\ Depth - 0.12835\ OM - 0.85848\ BD \cdot OM$
OLSK(P)	$logKsat = 2.11771 + 1.38330\ BD - 0.28621\ Slope - 15.28815\ Depth$
OLSK(R)	$logKsat = 0.07337 - 0.85962\ BD + 1.40465\ Slope - 0.73158\ Depth$

Slope was significant in the models predicting water retention at pF1.5, pF3.4 and $logK_{sat}$. In the PTF for water retention at pF3.4, the interaction term between slope and OM was significant. Surprisingly, slope was significant in the model predicting water content at pF1.5, even though the correlation between the two was almost zero.

4.2.2 Random forest models

The black box random forest models are more difficult to interpret that the OLS models, but measures of variable importance can help us understand what goes on in the model building. One approach is to measure which variables that on average reduce the residual error the most when chosen for a regression tree node split, or the *mean decrease in node impurity* (section 2.3.2.1 or James et al. (2013) for theory). Variance importance plots are presented for all RF models in figure 8.

As we can see from the figure, the random forest models follow the same patterns of variable importance as the linear models; bulk density is by far the most important variable for determining water retention in the low suction range, and is still important for higher suctions as well as for AWC and K_{sat}. The first three models for water retention close to saturation have pretty similar variable importance patterns. Slope and soil depth are slightly more important for water content at pF1.5 than for conditions closer to saturation, matching the variables included in the corresponding OLS PTF.

Compared to the low suction water retention models, vegetation seems to have more relative importance for predicting water retention at field capacity and higher suctions, as well as $logK_{sat}$. However, depth and slope are still more important in all models. Soil type is not very important in any model, but it had some importance in the RF4.2 and RFAW model, coinciding with the very different trends of the water retention curve and thus the available water, as seen in the boxplots in figure 7. Soil type was not even included in the $logK_{sat}$ model because of close to zero importance. Organic matter is the dominant explanatory variable for water retention at pF3.4, and is also important for the wilting point, together with bulk density.

4.3 PTF fits and test performances

The fit of all the models on the training set and their predictive performance on the test set are presented in table 4 by *train*- and *test* RMSE. In the cases where the OLS models include dummy variables creating separate functions, group-specific RMSEs were calculated *(Split train RMSE* and *Split test RMSE in the table*) to evaluate and compare the prediction error for either Histosols (H), soils from pine plantations (P) or other soils (R). The RMSE for the models' predictions on the whole test set is given in the table as the *total test RMSE*).

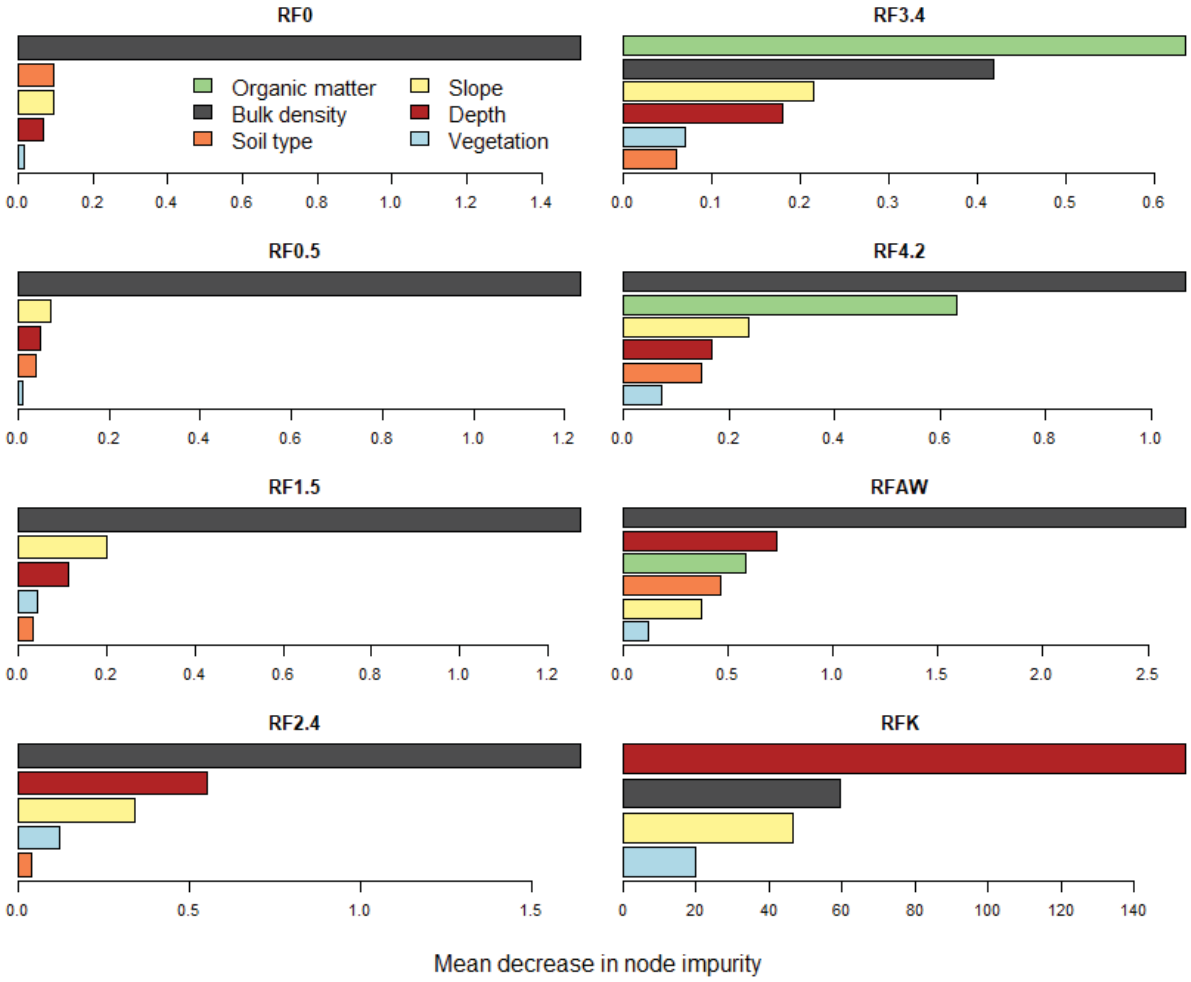

Figure 8. Variable importance plots for the random forest models, based on mean decrease in node impurity

Generally, the OLS models had a better fit than the RF models on the *training set*. The only exceptions are water retention at pFs 2.4, 3.4, 4.2 and AWC for pine plantation soils, and logKsat for other soils. The training set RMSEs were generally low for the pine plantation soils. While the OLS models fitted the training set best, the RF predicted better than the OLS models on the total test set and also the groups in the test set, for all of the predicted hydraulic properties. All the RF models' test error is lower than the training error, while the OLS models have lower training error than test error in around half the cases.

Predictions of water retention were especially accurate in the low suction range. As suction increases, predictions decrease in accuracy. Even though other variables like OM content, slope, depth and vegetation cover were more important in these functions, the obtained test RMSEs prove that there is still a lot of unexplained variance in the predictions. The greatest differences in predictive performance between the two modelling approaches, was for the models predicting Histosol water retention at field capacity, wilting point and AWC. RFK test error was much lower than the OLSK test error, even though model variables were the same.

A graphical comparison of the test performances of the two modelling methods can be seen in figure 9, where the observed water contents from the test set are plotted against the predicted values. Point predictions from all the models are presented in the same figure, and the symbol indicates the model. Both models have close fits to the 1:1 line between predicted and observed water contents. However, the random forest models clearly gave the most accurate predictions. From the figure it would seem that the water content at pF3.4 was the most difficult to predict for both the RF and the OLS models. This is confirmed by the high total test RMSEs of 0.0721 and 0.0760 respectively. Predictions of water retention at wilting point are more accurate, but RMSEs for water content at lower suctions are less variable. Predictions were more accurate on soils from Soldados and Tutupali than soils from Zhurucay.

The test set prediction accuracies of AWC models and logKsat models are compared graphically in figure 10. AWC predictions were more scattered in the Zhurucay soils, while they were more concentrated and accurate for the Soldados and Tutupali soils. Again, the RF model has the tightest fit, but the RMSEs were higher than 0.05 for both the OLS and RF model. The OLS model for predicting $logK_{sat}$ behaved almost like a class PTF, with predictions grouped together. The RF model has the better fit in this case as well, with only a few very inaccurate predictions.

Table 3. Comparison of model fit and test performances for both methods. H.: Histosol; P: Pine vegetation cover; R: not Histosol nor pine vegetation cover.

		N		PTF code		Split train RMSE		Total test RMSE		Split test RMSE	
		train	*test*	*OLS*	*RF*	*OLS*	*RF*	*OLS*	*RF*	*OLS*	*RF*
θ_{pF0}		260	85	OLS0	RF0	0.0324	0.0351	0.0333	0.0302	0.0333	0.0302
$\theta_{pF0.5}$		221	72	OLS0.5	RF0.5	0.0267	0.0275	0.0277	0.0241	0.0277	0.0241
$\theta_{pF1.5}$	H	20	5	OLS1.5(H)	RF1.5	0.0634	0.0714	0.0361	0.0348	0.0502	0.0478
	R	211	71	OLS1.5(R)		0.0391	0.0391			0.0349	0.0337
$\theta_{pF2.4}$	H	29	6	OLS.2.4(H)	RF2.4	0.1279	0.1381	0.0581	0.0481	0.1151	0.0654
	P	131	43	OLS2.4(P)		0.0538	0.0428			0.0478	0.0417
	R	100	36	OLS2.4(R)		0.0511	0.0527			0.0552	0.0518
$\theta_{pF3.4}$	P	87	42	OLS3.4(P)	RF3.4	0.0371	0.0335	0.0760	0.0721	0.0400	0.0340
	R	120	26	OLS3.4(R)		0.1022	0.1099			0.1120	0.1082
$\theta_{pF4.2}$	H	29	5	OLS4.2(H)	RF4.2	0.0865	0.1108	0.0575	0.0557	0.0641	0.0478
	P	129	43	OLS4.2(P)		0.0419	0.0368			0.0428	0.0427
	R	79	27	OLS4.2(R)		0.0984	0.1080			0.0743	0.0727
AWC	H	29	5	OLSAW(H)	RFAW	0.1475	0.1621	0.0747	0.0642	0.1171	0.0636
	P	129	43	OLSAW(P)		0.0537	0.0506			0.0446	0.0429
	R	79	27	OLSAW(R)		0.0969	0.1056			0.0990	0.0882
$logK_{sat}$	P	131	43	OLSK(P)	RFK	0.7169	0.7225	0.9313	0.7535	0.8464	0.6701
	R	128	42	OLSK(R)		0.8882	0.8490			1.0109	0.8303

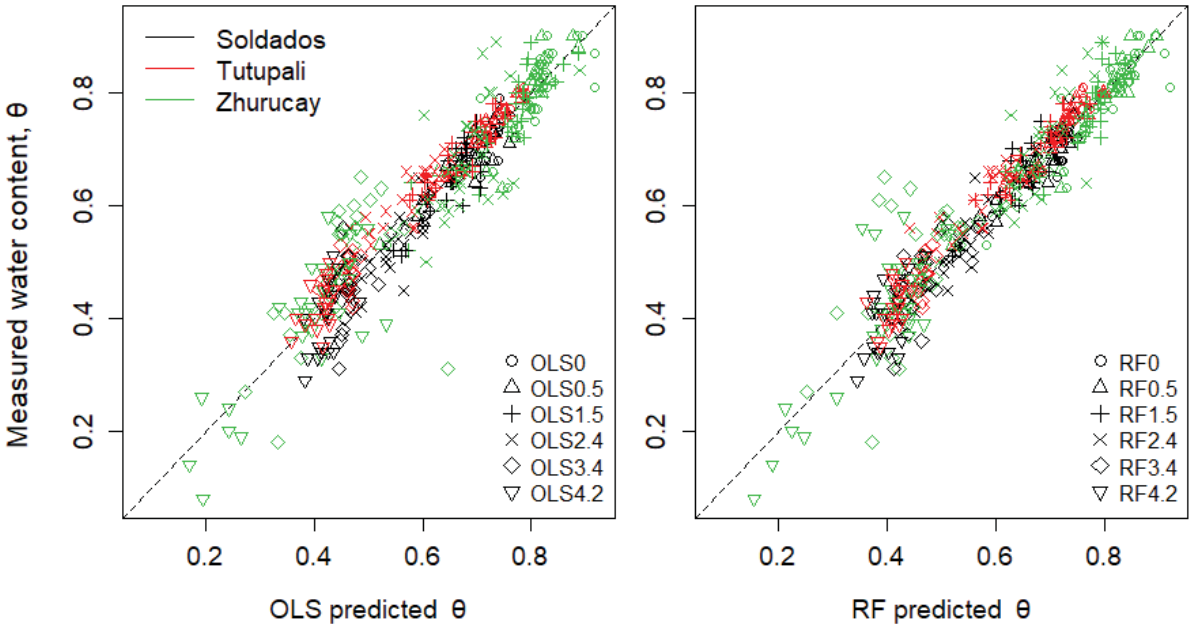

Figure 9. Predicted vs. measured values for water retention at all matric potentials evaluated in this thesis. Points are coloured after site. OLS models on the left and RF models on the right. Diagonal dotted line is the 1:1 relationship.

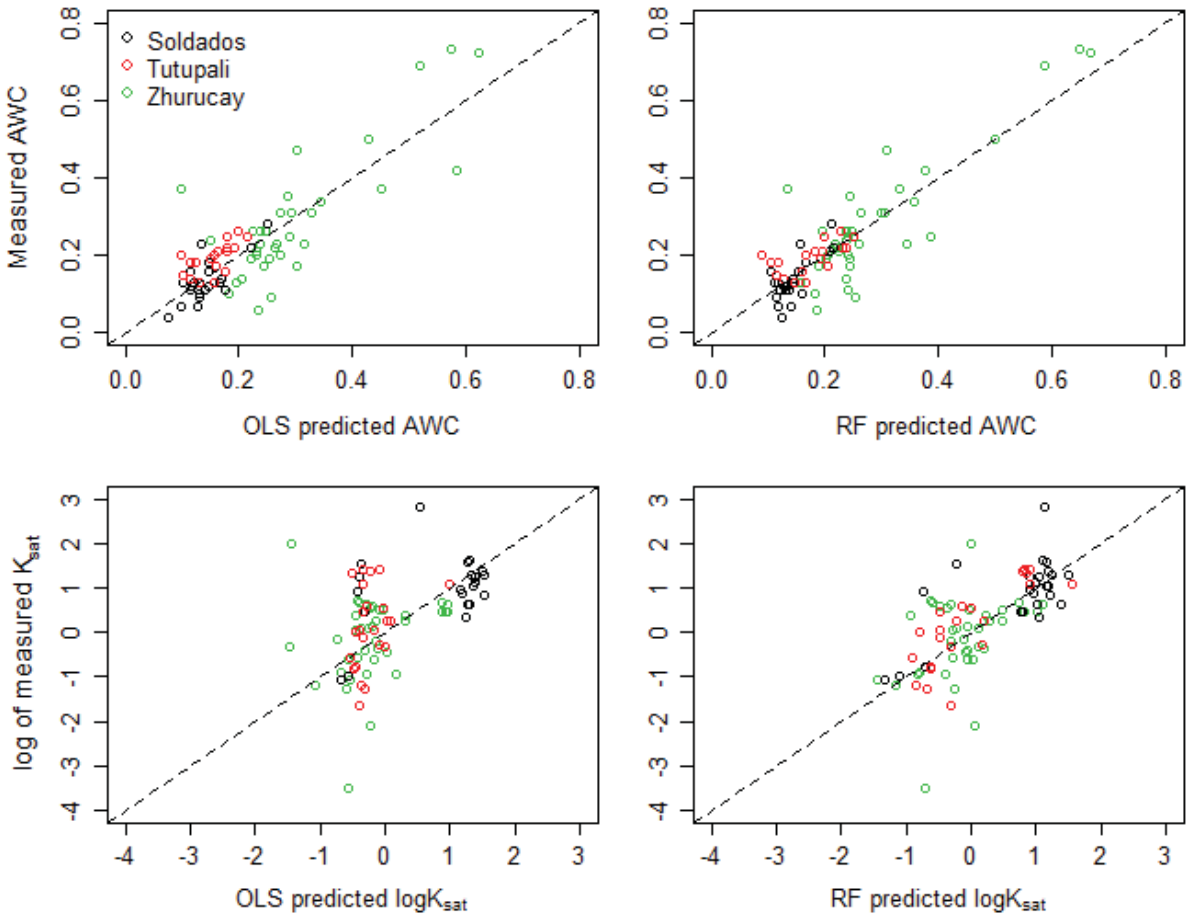

Figure 10. Predicted vs. measured available water capacity (AWC) and logKsat, coloured by site

4.3.1 Van Genuchten water retention curves

Figure 11 presents some examples of the fitted Van Genuchten curves on the test set, were the red points represent the measured water retention values and the blue and green points the estimates from the RF and OLS models respectively. The fitted functions from the estimates were close to the red line fitted from the measured data in most cases from all the four projects. Sometimes it was necessary to adjust initial values to fit a curve and other times it was not possible to fit a curve, even with adjusted initial values. Nine of the observations from the IC project had only three available water retention points which is fewer than the amount needed to fit a curve from the VG expression that has four unknown parameters. Where organic matter data was missing, no predictions of water retention at pFs 3.4 and 4.2 were available, which gave convergence problems as well. Even on the six observations that lacked only one water retention data point, the curve was not fitted. Of the 62 rows in the data set where enough water retention points were present, both observed and predicted, three fitted curves were obtained in 49 of the cases.

The RMSD was calculated using only observations from the rows in the test set were all the three curves were fitted successfully. Results showed that the distance from the red curve to the curves fitted from the WRC predictions of the two modelling methods were quite similar; the OLS had an RMSD of 0.0332, while the RF had a slightly better RMSD of 0.0305

4.4 Borja PTF performance on test set

The predictive performance of the selected functions developed by Borja (2006) on the test set of this thesis is presented in table 5 and the corresponding RMSEs of the PTFs developed in this thesis is shown in the adjacent grey table. Borja's PTFs obtained a smaller test error with the data in this thesis than in the original publication (table 2). In the low suction range, Borja's PTFs did not do a better job at predicting the water content than the PTFs developed in this thesis. However, Borja's functions predicted mostly better than the PTFs of this thesis in the high suction range.

Table 4. Prediction error of the PTFs developed by Borja (2006) on the test set of this thesis. H: Histosols; P: pine plantation soils; R: not Histosols nor pine plantation soils. Grey table on the right shows RMSEs of corresponding PTFs developed in this thesis.

	Code		N	Total *test* RMSE	Split *test* RMSE
θ_{pF0}	MRLM2a		85	0.0418	0.0418
$\theta_{pF1.5}$	MRLM2b	H	6	0.0378	0.0487
		R	71		0.0368
$\theta_{pF3.4}$	MRLM4e	P	43	0.0424	0.0338
		R	41		0.0565
$\theta_{pF4.2}$	MRLM4f	H	5	0.0466	0.0860
		P	42		0.0402
		R	29		0.0482

Split RMSE OLS	*Split RMSE RF*
0.0333	*0.0302*
0.0502	*0.0478*
0.0349	*0.0337*
0.0400	*0.0340*
0.1120	*0.1082*
0.0641	*0.0478*
0.0428	*0.0427*
0.0743	*0.0727*

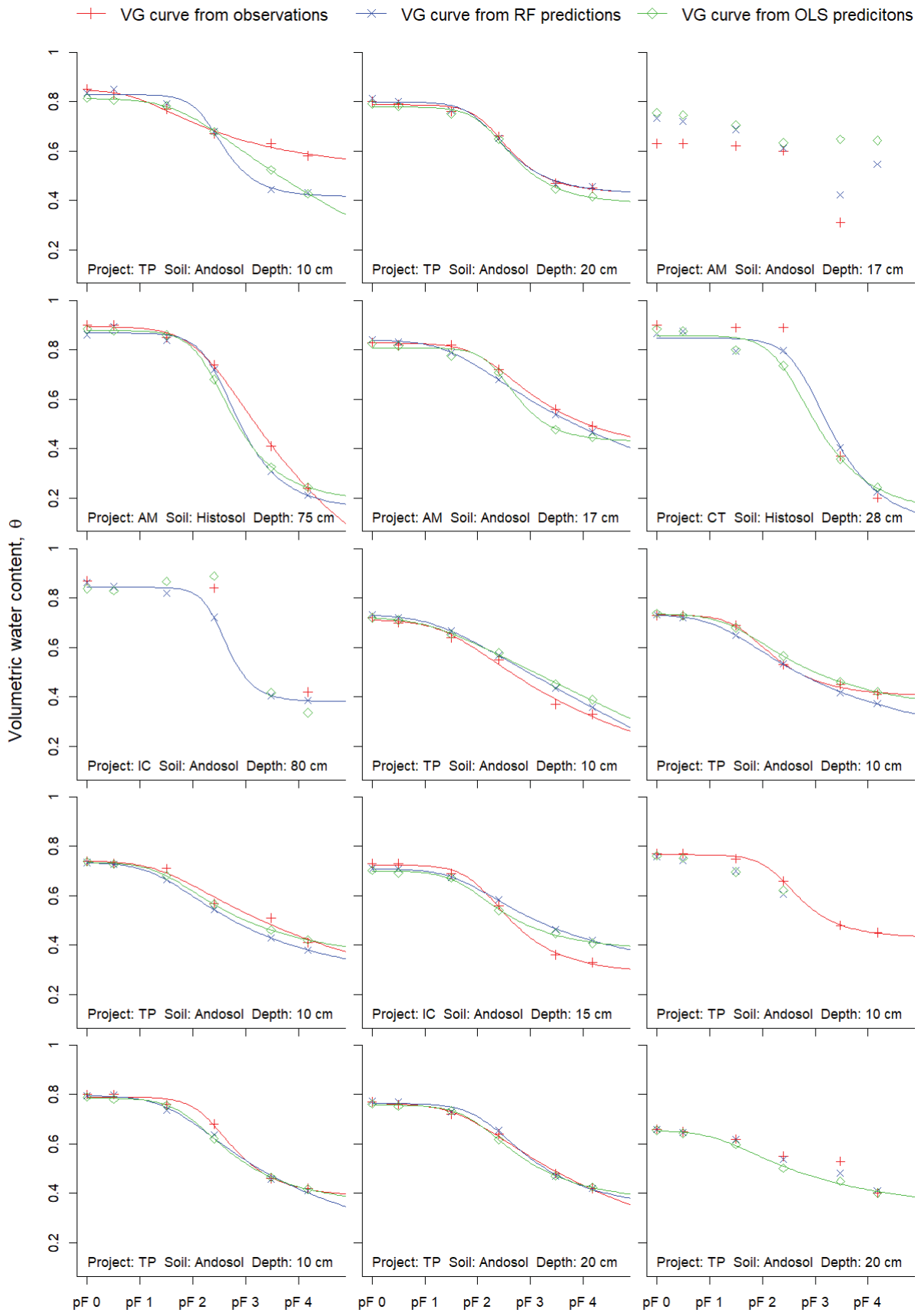

Figure 11. Fitted Van Genuchten curves on fifteen selected rows of the test set. Red, blue and green points represent measured values, RF estimates and OLS estimates respectively, and the continuous lines the VG curves fitted from the points.

5 Discussion

5.1 Selected predictors and their interactions with soil hydraulic properties

In the models developed in this thesis, *bulk density* was by far the most important variable for predicting water retention in both the low and the high suction range. BD is a strong indicator of the soil structure and pore size distribution, properties that determine water retention at matric potentials close to saturation. Soil *organic matter* content basically tells the same story, and it did not provide enough additional information to be included in the low-suction PTFs. A soil's pore sizes are reduced as OM decomposes, hence the pore size distribution of equally organic soils can be quite different (Rezanezhad et al., 2016). Grover and Baldock (2013) studied the relationship between peat hydrology and its chemical composition. They measured the proportions of *carbohydrates*, *protein*, *lignin*, *lipids*, *carbonyl* and *char* in the peat, which change as the OM decomposes, and linked it to K_{sat} and soil water retention in the low suction range. Their results showed that the chemical composition of the OM explained much more of the variation in hydraulic properties than bulk density. In this thesis however, BD was the best of the available explanatory variables at indicating OM chemical composition from decomposition, which explains its importance in both the water retention models and the K_{sat} models. OLSK(P) is the only regression model that has included a positive BD term, which is unexpected. Some of the pine observations had very high saturated hydraulic conductivity compared to the rest of the dataset, probably because of preferential flow along roots, which may have had an effect in the model.

The importance of OM content on water retention at high suctions is somewhat discussed in the literature (Rawls et al., 2004). OM is known to compete with clay minerals and thus reduce water retention in fine-textured soils (Christensen, 1996). However, Rawls et al. (2003) conclude in their study that the effect of organic carbon on water retention is always positive for soils with an organic carbon higher than 5%, i.e. an OM content higher than 10%, using the conversion factor suggested by Pribyl (2010). This claim does not hold for the extremely organic soils studied in this thesis with a mean OM content of 33%. In the scatterplot in figure 6 we see that there is threshold at approximately 50% OM content where water retention in the high suction range goes from a positive correlation with OM to being negatively affected by an increase in OM content. However, the negative trend is very scattered, and this is where the interaction between BD and OM becomes important. Organic soils that contain high amounts of poorly decomposed OM have less specific surface for adsorption than soils consisting of more decomposed or compressed OM, and water retention at high suctions will increase as BD increases. A positive OM:BD interaction term is included in the high suction OLS models, making sure that the originally negative effect of

increasing BD becomes positive in highly organic soil. The interaction term is negative for in the OLSAW model, since a higher water retention at wilting point means less available water.

The importance of *soil depth* in the models OLS1.5, OLS2.4, OLSAW and OLSK, is probably related to the vertical variation of the degree of compaction and OM decomposition affecting pore space distribution. Hoag and Price (1995) found that K_{sat} in the peat of their study dropped by several orders of magnitude from the surface to deeper, more humified layers. While soil depth is relatively important in the RF models at high suctions (figure 8), it has not been included in the corresponding OLS models. This is probably due to the OM:BD interaction explaining most of the information on the degree of OM decomposition relevant in this suction range. Most Andosol samples were taken in the top 25 cm of the soil, while Histosols dominate the deepest soil samples in the database, and have been separated out with dummy variables in the OLS models that include soil depth. Care must be taken to avoid extrapolation and using the PTFs developed in this thesis to predict hydraulic properties in deep layers of Andosols.

Slope is significant in the OLS1.5, OLS3.4 and OLSK models, and relatively important in RF models for higher suctions than saturation. Topography provides useful information about where fine soil particles are more likely to be transported from and to by erosion and also about the drainage conditions deciding OM decomposition and weathering of the soil (Romano & Chirico, 2004). The slope effect is negative in the OLS3.4 model, where micropores and specific surface is important for water retention, and slightly positive in the OLS1.5 model, where large pores are important. The effect in the logKsat model for the natural soils is positive. This indicates that pore space and -interconnectivity increase with the steepness of the slope. When slope is close to zero, water retention is mainly explained by the other model variables. The Histosol dummy variable included in OLS1.5 and the slope:OM interaction term included in the OLS3.4 model both account for differences in OM accumulation between slope bottoms with almost water saturated conditions and dryer hill tops, which may have the same slope gradient.

According to the RF variance importance plots in figure 8, the categorical variables *soil type* and *vegetation cover* were not particularly useful for explaining the soil water dynamics in the study area. However, the *Pine* and *Histosol* dummy variable were important enough to include in many of the OLS water retention models. Pine plantation soils had less organic matter than the average, and Histosols had more than the average. Using the dummy variables is an efficient way of discriminating these different OM levels in the linear regression approach. Practically all soils belonging to the vegetation group *moss and cushion plants*, coincide with the Histosol observations, thus this category did not provide any more information on soil water dynamics. Vegetation was most

important for the discrimination of pine tree soils in the K_{sat} models. Jorda et al. (2015) studied hydraulic conductivity in areas of natural vegetation and agricultural land with different tillage practices and found that bulk density and land use were the most important factors in determining K_{sat}. There was no arable land in the study area of this thesis, but the pine plantation soils had, on average, higher saturated hydraulic conductivity than the other soils. This may be explained by preferential flow along the roots of pine trees, and it may also be attributed to differences in the OM composition developed during the last 20 years since the establishment of the pine plantations.

5.2 PTF fits and test performances

While the linear regression models generally had lower train RMSE than the random forest models, training set variation in high suction water retention for pine plantation soils, and logKsat for other soils, was better explained by the RF models. The mean organic matter content of the pine plantation observations in the training set was 23%, thus soil texture is probably a more influencing factor for water retention at high suctions in these soils. As soil texture was not included in the model development, other variables that correlate to soil texture and their interactions become important. Approaches based on regression trees and other data mining techniques, have an extraordinary ability to detect non-linear structures in the data and utilise all available information (Pachepsky & Schaap, 2004), while it would seem that the OLS method is not as good at capturing the complex relationships between texture-dependent hydraulic properties and the available predictors in the case of the less organic soils. The train RMSE was generally low for the pine plantation soils, which could indicate a biased model. However, we see from the boxplot in figure 7 that these were also the soils with least variation in water retention, thus a tighter model fit is to be expected. Saturated hydraulic conductivity is also linked to soil texture, but only because texture is an indicator of the interconnected pore space of the soil. It does not depend on the soils adsorptive strength like water retention in the high suction range. While the OLSK has a lower train RMSE for the pine plantation soils, the RFK model does a better job at capturing the relationship between interconnected pore space and the available predictors in the highly variable soils with natural vegetation cover.

A look at the models' *test* RMSE paints a different picture of the two modelling approaches. The RF models predicted better than OLS models in every case. This indicates that the OLS models are slightly overfitted, a problem the random forest approach theoretically does not have, due to the *Law of Large Numbers* (Breiman, 2001; James et al., 2013). The fact that the test error is lower than the training error for the majority of RF models, proves that they have accomplished a good data generalization. The OLS models have lower training error than test error in around half the cases, which is still pretty good, as the training error is normally lower than test error when predicting on a dataset that has not been used in the model development. Total test RMSEs of all the

developed water retention models are well inside the RMSE range of 0.02 – 0.11 m^3 m^{-3} reported in the literature (Donatelli et al., 2004; Wösten et al., 2001).

Generally, the test error of the OLS models is not much higher than the corresponding RF model test error. Some exceptions are the test RMSEs of the models predicting Histosol water retention at field capacity, wilting point and AWC. The large RMSE differences could be random, as the test set has very few Histosol observations. However, it could also be attributed to slope not being included in the OLS functions predicting these water retention points. From the variable importance plots in figure 8, we see that slope is relatively important for field capacity, wilting point and AWC, and this importance is captured in the RF models. The test error difference between OLSK and RWK was also considerable, even though the included model variables were the same. Depth was the most important variable in the logKsat models, thus the clustering of OLSK predictions seen in figure 10 is probably due to the fact that most of the observation were sampled at soil depths 10 and 20 cm. There are complex interactions between soil properties and environmental factors that determine the permeability of a soil and thus its saturated hydraulic conductivity. By choosing between a limited set of *m* random predictors of the total set of predictors at each node split, the RF models are better at capturing all information available in the data set.

Predictions of water retention were especially accurate in the low suction range, were bulk density alone explained most of the variance. As soil-water suction increases, and the soil's adsorptive properties determines water retention, the lack of soil texture and OM composition as explanatory variables in the PTFs is reflected in less accurate predictions, both for OLS and RF models. Test RMSEs for the less organic pine plantation soils were below 0.05 for all water retention models, while the other more organic soils had high RMSEs in the high suction range. This could indicate that textural variation is well explained by other predictors and that OM composition is mainly responsible for the variation in soil hydraulic properties. It would also explain why the models predict K_{sat} better in the pine plantation soils than in the other soils. Moreover, Hemond and Goldman (1985) argue that the auger-hole methods used to determine field K_{sat} in this thesis, do not give reliable measurements in compressible soils, due to alterations of the soil matrix near the auger hole, or compression by the people making the measurements. If this is the case for the soils of the study area, measurements from the very organic grassland or peat soils would not be as reliable as the less organic pine plantation soils. Measurement error would affect both train and test RMSE. Most of the pine plantation soils were sampled in Tutupali and Soldados, hence the tight fit of these observations in the predicted vs. measured plots in figures 9 and 10. The majority of the grassland data and all the peat data were from Zhurucay, explaining why the green points are more scattered in the plots. This was also the only site that was studied in all the four projects

providing the database of this thesis, and systematic differences in methodology may have led to more variance.

5.2.1 van Genuchten water retention curves

VG curves were fitted successfully on both measured and predicted water retention points in 49 of the 62 cases in the test set where enough data was available. Of the 13 remaining cases that had enough water retention data, but did not manage to converge all three curves, five were from the TP project (one from Soldados and four from Tutupali) and the last eight were from the AM project in Zhurucay. Convergence was achieved with the observed data in only three of the 13 cases, and on the predicted data from both methods in five of the cases. The convergence problems are possibly attributed to measurement or prediction error, but it could also be explained by inapplicability of the VG expression in describing the very unique water retention properties of Andosols and Histosols (Dettmann et al., 2014; Vereecken et al., 2010).

The root mean squared differences between the curves fitted on the predicted data and the corresponding curve fitted on observed data was very low, with RMSDs below 0.035 for obtained for both the OLS and RF curve fits. In their study on Brazilian soils, Tomasella et al. (2003) achieved an RMSD of 0.087. Ahuja et al. (1985) obtained an RMSD of 0.05. However, the results from this thesis might be a little misleading, as only the 49 rows in the data sets where all curves were fitted, were used in the RMSD calculation. If a curve were to be adjusted to best describe the remaining problematic cases, mean RMSD for the whole test set would most probably increase. Regardless of the true RMSD of the test set, the RMSDs obtained from the 49 cases are useful for comparing the two modelling methods. It would seem that, provided that convergence is achieved, both the OLS method and the RF method are useful for obtaining relatively accurate VG equation parameters through PTF predictions of water retention at pFs 0, 0.5, 1.5, 2.4, 3.4 and 4.2.

5.3 Borja PTF evaluation and performance on test set

In his PTF development in 2006, Borja included some samples from northern Ecuador in addition to soils close to the study area of this thesis. The Andosols in northern Ecuador are younger, coarser and contain less organic matter, thus the hydraulic properties are naturally different. In addition, the northern soils of his study were more affected by agricultural activity. He found that the accuracy of his PTFs was higher in the soils from the south and that predictions of water contents at matric potentials close to saturation tended to be overestimated in the northern soils. This is probably the reason why the prediction errors obtained with the test set of this thesis were so much lower than the corresponding errors of the original publication (table 2).

Even though Borja's PTFs were developed using only Andosol data, the MRLM2b model including OM and θ_{pF0} as predictor variables, predicted better than the OLS1.5 model on the Histosol observations. However, the PTFs developed in this thesis of did a better job at predicting Andosol water content at pF1.5 and water content at saturation. The MRLM4f model including $\theta_{pF3.4}$ as a predictor variable did not predict Histosol wilting point water retention as well as the PTFs of this thesis, and the test RMSEs for high suction water retention in pine plantation soils were similar in all three cases. However, Borja's high suction PTFs predicted much more accurately on Andosols with natural vegetation cover. Hence, Borja's functions can be used to improve prediction accuracy in these soils. Borja also developed functions using artificial neural network analysis, and the model predicting water retention at pF3.48 performed even better than the MRLM4e function in the original publication (table 2). If this model is available, it may be an even better choice for improving prediction.

Avoiding the cost and impracticalities of measuring water retention is the main idea behind PTF development, thus the necessity of improved prediction has to be assessed before using Borja's functions. However, analysis of high suction water retention points does not require bringing steel ring equipment out in the field for sampling undisturbed samples, and it is not especially time consuming. If samples are taken anyway to analyse soil BD and OM for the PTF inputs, it might not end up too costly doing a parallel pressure chamber analysis of one high suction water retention point, if the necessary equipment is already available.

5.4 Study limitations

The number of peat observations in the dataset used for the development of the PTFs in this thesis was very limited, with 29 and 6 Histosols represented in the training- and test set respectively. It was not ideal that all Histosol observations were from only one of the projects at one site. In addition, 19 of the 35 Histosol observations were taken from soil depths of 40 cm or deeper, and some Histosol samples are from the C horizon which is much lower in organic matter. Hence, a reliable picture of peat soil water dynamics in the Ecuadorian páramo might not be captured in the developed PTFs.

Another limitation of this study is the difference in methodology between the different projects. While the authors in the AM, IC and TP projects used the sandbox and pressure chamber methods to analyse water retention, the authors of the CT project used the *multistep outflow* method, which may have given systematic differences in the measurements. The procedures of organic matter analysis were not the same between the methods. While organic matter content in the AM, CT and IC projects was determined from samples taken in the same soil profile as the ring samples, organic matter in the many pine plantation observations in the TP project was analysed on only one sample

in each of the two depths for every 24x24 blocks. This means that there are three different sets of water retention observations from each block that all are linked to the same organic matter content. Most of the organic matter observations were conversions from organic carbon contents, while organic matter in the AM project was measured directly. I proposed a carbon to OM conversion factor of 2 instead of the classic 1.72, which may also have had an effect on the results of this thesis.

The number of observations in the different projects and at the different sites were not the same, and the fact that random effects were excluded in the PTF development may have led to a suboptimal generalization of the data and possible bias in the functions' predictions. The reliability of the functions on other data sets remains to be studied in future research in the area.

6 Future research

To improve pedotransfer functions for predicting hydraulic properties of highly organic páramo soils, information on the degree of OM decomposition is recommended. Either using the simple and well-known Von Post scale (von Post & Granlund, 1926) in the field, or by measuring the chemical composition of the OM, as done very successfully by Grover and Baldock (2013).

The mineralogy of the soil is also known to affect its structure and pore size distribution (Bruand, 2004; K. & J., 2005), and thus the water retention in the low suction range and K_{sat}. The way in which clay minerals are arranged in the soil, the minerals' size and surface charges are variables that become increasingly important for high suction water retention (Quirk, 1994). In the case of Andosols, the interaction between pH and Andosol mineralogy could be a significant explanatory variable (Shoji et al., 1996). Buytaert et al. (2005a) showed that the properties of Ecuadorian soils linked to the volcanic ash content vary with the distance to the still active volcanoes Tungurahua and Sangay in central Ecuador. Hence, a geostatistical modelling approach may be appropriate to improve the prediction of more texture dependent soil hydraulic variables in the páramo of southern Ecuador.

To address the reliability of the PTFs, it is important to evaluate their predictive performance using other databases than the ones used for the PTF development. If they turn out to be reliable for similar soils, the pedotransfer functions developed in this thesis can be a useful tool for hydrological modelling on a large scale in the very vulnerable, important and precious páramo of southern Ecuador.

7 Conclusions and recommendations

The purpose of this thesis was to contribute to the understanding of the variables affecting soil hydrology in the Ecuadorian páramo ecosystem. This was done by developing pedotransfer functions to predict points on the water retention curve and saturated hydraulic conductivity for non-allophanic Andosols and Histosols. Soil texture was not included in the analysis, but *bulk density*, *soil organic matter*, *slope*, *soil depth*, *vegetation cover* and *soil type* were used as explanatory variables in the development of both linear regression models and random forest models. Bulk density was by far the most important variable for predicting soil hydraulic properties in the study area. Organic matter content and its interaction with bulk density was important at high soil-water suctions. Soil depth was the most important variable for the saturated hydraulic conductivity models.

The random forest models obtained lower prediction error (RMSE) than the corresponding linear regression models for all the studied soil hydraulic properties. Linear regression predictions of saturated hydraulic conductivity had a high RMSE of 0.9313, while the corresponding random forest performed better, with a RMSE of 0.7535. Predictions of water retention in the low soil-water suction range were satisfactory; both the linear regression models and the random forest models obtained RMSE values below 0.05 for water contents at suctions lower than field capacity. Model predictions of water retention at higher soil-water suctions gave RMSE values between 0.05 and 0.08. The random forest models had stronger predictive power than the linear regression models in this thesis, but the difference between the two approaches was not dramatic. When fitting the van Genuchten curve on the predicted water retention points, both approaches gave good generalizations of soil moisture characteristics. In many practical situations, the simpler, more interpretable linear regression model will often be preferred.

For better predictions of water retention in the high suction range, the functions developed by Borja Ramón that include water retention information as explanatory variables can be used, or new functions can be developed that include other explanatory variables. Soils in the páramo of southern Ecuador are highly organic, hence the chemical composition of the organic matter is a viable predictor candidate for further improvement of PTFs in the area. The incorporation of soil mineralogy coupled with pH and/or geostatistical analysis, could also yield better results.

8 References

Ahuja, L., Naney, J. & Williams, R. (1985). Estimating Soil Water Characteristics from Simpler Properties or Limited Data 1. *Soil Science Society of America Journal*, 49 (5): 1100-1105.

Akpa, S. I. C., Ugbaje, S. U., Bishop, T. F. A., Odeh, I. O. A. & Varennes, A. (2016). Enhancing pedotransfer functions with environmental data for estimating bulk density and effective cation exchange capacity in a data-sparse situation. *Soil Use and Management*, 32 (4): 644-658. doi: doi:10.1111/sum.12310.

Aucapiña, G. A. & Marín, F. G. (2014). *Efectos de la posición fisiográfica en las propiedades hidrofísicas de los suelos de páramo de la microcuenca del Río Zhurucay*. Cuenca: University of Cuenca.

Batjes, N. H. (1996). Development of a world data set of soil water retention properties using pedotransfer rules. *Geoderma*, 71 (1): 31-52. doi: https://doi.org/10.1016/0016-7061(95)00089-5.

Borja, P. M. (2006). *Desarollo de Funciones de Edafo-Transferencia para la Caracterizacion Hidraulica de Andosoles*. Cuenca: Univeristy of Cuenca.

Bouma, J. & Lanen, H. A. J. v. (1987). *Transfer functions and threshold values: from soil characteristics to land qualities*. Proceedings of the international workshop on Quantified land evaluation procedures : held in Washington, DC, 27 April - 2 May 1986.

Bouma, J. (1989). Using Soil Survey Data for Quantitative Land Evaluation. In Stewart, B. A. (ed.) vol. 9 *Advances in Soil Science*, pp. 177-213. New York: Springer.

Brady, N. C. & Weil, R. (2014). *Elements of the Nature and Properties of Soils*. 3 ed. Harlow: Pearson Education Limited.

Brahim, B. H. (1987). *Influence des constituants alumineux et feniques non cristallins sur les cycles du carone et de l'agote dans les sols montegands acids*: Université de Nancy I.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45 (1): 5-32. doi: 10.1023/a:1010933404324.

Briggs, L. J. & McLane, J. W. (1907). *The moisture equivalents of soils*. U.S. Government Printing Office: (U.S. Department of Agriculture Bureau of Soils. Bulletin).

Brooks, R. & Corey, A. (1964). Hydraulic properties of porous media. *Hydrology Papers 3, Colorado State Univ., Fort Collins.*

Bruand, A. (2004). Utilizing mineralogical and chemical information in PTFs. In Development of Pedotransfer Functions in Soil Hydrology, vol. Volume 30 *Developments in Soil Science*, pp. 153-158: Elsevier.

Buurman, P., de Boer, K. & Pape, T. (1997). Laser diffraction grain-size characteristics of Andisols in perhumid Costa Rica: the aggregate size of allophane. *Geoderma*, 78 (1): 71-91. doi: https://doi.org/10.1016/S0016-7061(97)00012-8.

Buytaert, W., Deckers, J., Dercon, G., de Bièvre, B., Poesen, J. & Govers, G. (2002). Impact of land use changes on the hydrological properties of volcanic ash soils in South Ecuador. *Soil Use and Management*, 18 (2): 94-100. doi: 10.1111/j.1475-2743.2002.tb00226.x.

Buytaert, W., De Bievre, B., Wyseure, G. & Deckers, J. (2004). The use of the linear reservoir concept to quantify the impact of changes in land use on the hydrology of catchments in the Andes. *Hydrology and Earth System Sciences*, 8 (1): 108-114.

Buytaert, W., Sevink, J., De Leeuw, B. & Deckers, J. (2005a). Clay mineralogy of the soils in the south Ecuadorian páramo region. *Geoderma*, 127 (1): 114-129. doi: https://doi.org/10.1016/j.geoderma.2004.11.021.

Buytaert, W., Wyseure, G., De Bièvre, B. & Deckers, J. (2005b). The effect of land-use changes on the hydrological behaviour of Histic Andosols in south Ecuador. *Hydrological Processes*, 19 (20): 3985-3997. doi: 10.1002/hyp.5867.

Cajamarca, J. & Tenorio, G. (2008). *Estudio Geomorfológico y de suelos del páramo de Quimsacocha*. Cuenca: University of Cuenca.

Cárdenas, I. L. (2014). *Impacto de las prácticas agropecuarias y forestales sobre las propiedades físico-químicas de los suelos andinos del sur del Ecuador*. Cuenca: University of Cuenca.

Chen, C. & Payne, W. (2001). Measured and modeled unsaturated hydraulic conductivity of a Walla Walla silt loam. *Soil Science Society of America Journal*, 65 (5): 1385-1391.

Christensen, B. (1996). Carbon in primary and secondary organomineral complexes. In vol. 27 *Structure and organic matter storage in agricultural soils*, pp. 97-165: CRC Press Boca Raton, FL.

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). *Journal of Hydrology*, 407 (1): 164-174. doi: http://dx.doi.org/10.1016/j.jhydrol.2011.07.021.

Dettmann, U., Bechtold, M., Frahm, E. & Tiemeyer, B. (2014). On the applicability of unimodal and bimodal van Genuchten–Mualem based models to peat and other organic soils under evaporation conditions. *Journal of Hydrology*, 515 (Supplement C): 103-115. doi: https://doi.org/10.1016/j.jhydrol.2014.04.047.

Donatelli, M., Wösten, J. H. M. & Belocchi, G. (2004). Methods to evaluate pedotransfer functions. In Development of Pedotransfer Functions in Soil Hydrology, vol. Volume 30 *Developments in Soil Science*, pp. 357-411: Elsevier.

Durner, W. & Peters, A. (2009). *SHYPFIT 2.0 – Software zur Anpassung hydraulischer Funktionen an Messdaten*.

FAO. (2006). *Guidelines for soil description*. Rome: Food and Agriculture Organization of the United Nations.

FAO. (2009). *Guía para la descripción de suelos*. Roma: Food and Agriculture Organization of the United Nations.

G. Schaap, M., J. Leij, F. & Van Genuchten, M. (2001). ROSETTA: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. *J Hydrol.*, 251: 163-176. doi: 10.1016/S0022-1694(01)00466-8.

Givi, J., Prasher, S. O. & Patel, R. M. (2004). Evaluation of pedotransfer functions in predicting the soil water contents at field capacity and wilting point. *Agricultural Water Management*, 70 (2): 83-96. doi: https://doi.org/10.1016/j.agwat.2004.06.009.

Grover, S. P. P. & Baldock, J. A. (2013). The link between peat hydrology and decomposition: Beyond von Post. *Journal of Hydrology*, 479: 130-138. doi: https://doi.org/10.1016/j.jhydrol.2012.11.049.

Hemond, H. F. & Goldman, J. C. (1985). On Non-Darcian Water Flow in Peat. *Journal of Ecology*, 73 (2): 579-584. doi: 10.2307/2260495.

Hillel, D. (2003). *Introduction to environmental soil physics*: Academic press.

Hoag, R. S. & Price, J. S. (1995). A field-scale, natural gradient solute transport experiment in peat at a Newfoundland blanket bog. *Journal of Hydrology*, 172 (1): 171-184. doi: https://doi.org/10.1016/0022-1694(95)02696-M.

Hodnett, M. & Tomasella, J. (2002). Marked differences between van Genuchten soil water-retention parameters for temperate and tropical soils: A new water-retention pedo-transfer functions developed for tropical soils. *Geoderma*, 108: 155-180. doi: 10.1016/S0016-7061(02)00105-2.

Hofstede, R., Segarra, P. & Mena-Vásconez, P. (2003). *Los Páramos del Mundo. Proyecto Atlas Mundial de los Páramos.* Quito: Global Peatland Initiative/NC-IUCN/EcoCiencia.

Hungerbühler, D., Steinmann, M., Winkler, W., Seward, D., Egüez, A., Peterson, D. E., Helg, U. & Hammer, C. (2002). Neogene stratigraphy and Andean geodynamics of southern Ecuador. *Earth-Science Reviews*, 57 (1): 75-124. doi: http://dx.doi.org/10.1016/S0012-8252(01)00071-X.

IUSS Working Group WRB. (2006). *World reference base for soil resources 2006*. World Soil Resources Reports, 103. Rome: FAO.

IUSS Working Group WRB. (2015). *World Reference Base for Soil Resources 2014, update 2015*. World Soil Resources Reports, 106. Rome: FAO.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An introduction to statistical learning*, vol. 112: Springer.

Jorda, H., Bechtold, M., Jarvis, N. & Koestel, J. (2015). Using boosted regression trees to explore key factors controlling saturated and near-saturated hydraulic conductivity. *European Journal of Soil Science*, 66 (4): 744-756. doi: 10.1111/ejss.12249.

K., D. & J., S. (2005). Clay mineralogy determines the importance of biological versus abiotic processes for macroaggregate formation and stabilization. *European Journal of Soil Science*, 56 (4): 469-479. doi: doi:10.1111/j.1365-2389.2004.00682.x.

Karube, J. & Abe, Y. (1998). Water retention by colloidal allophane and imogolite with different charges. *Clays and clay minerals*, 46 (3): 322-329.

Koestel, J. & Jorda, H. (2014). What determines the strength of preferential transport in undisturbed soil under steady-state flow? *Geoderma*, 217-218: 144-160. doi: https://doi.org/10.1016/j.geoderma.2013.11.009.

Mizota, C. & Van Reeuwijk, L. (1989). Clay mineralogy and chemistry of soils formed in volcanic material in diverse climatic regions. *Clay mineralogy and chemistry of soils formed in volcanic material in diverse climatic regions.* (2).

Nanzyo, M., Shoji, S. & Dahlgren, R. (1993). Physical Characteristics of Volcanic Ash Soils. In Shoji, S., Nanzyo, M. & Dahlgren, R. (eds) vol. 21 *Developments in Soil Science*, pp. 189-207: Elsevier.

Nemes, A., Rawls, W. J. & Pachepsky, Y. A. (2006). Use of the Nonparametric Nearest Neighbor Approach to Estimate Soil Hydraulic Properties. *Soil Science Society of America Journal*, 70: 327-336. doi: 10.2136/sssaj2005.0128.

Nielsen, D. R. & Shaw, R. H. (1958). Estimation of the 15-atmosphere moisture percentage from hydrometer data. *Soil Science*, 86 (2): 103-105.

Oosterbaan, R. J. & Nijland, H. J. (1994). Determining hydraulic conductivity of soils. In Ritzema, H. P. (ed.) *Drainage principles and applications*. Wageningen, The Netherlands: International Institute for Land Reclamation and Improvement (ILRI).

Pachepsky, Y. & Schaap, M. G. (2004). Data mining and exploration techniques. In Development of Pedotransfer Functions in Soil Hydrology, vol. Volume 30 *Developments in Soil Science*, pp. 21-32: Elsevier.

Padrón, R., Wilcox, B., Crespo, P. & Célleri, R. (2015). Rainfall in the Andean Páramo—New Insights from High-Resolution Monitoring in Southern Ecuador. *Journal of Hydrometeorology*, 16. doi: 10.1175/JHM-D-14-0135.1.

Poulenard, J., Podwojewski, P., Janeau, J. L. & Collinet, J. (2001). Runoff and soil erosion under rainfall simulation of Andisols from the Ecuadorian Paramo: effect of tillage and burning. *Catena*, 45 (3): 185-207. doi: 10.1016/s0341-8162(01)00148-5.

Pribyl, D. W. (2010). A critical review of the conventional SOC to SOM conversion factor. *Geoderma*, 156 (3): 75-83. doi: https://doi.org/10.1016/j.geoderma.2010.02.003.

Quichimbo, P., Tenorio, G., Borja, P., Cárdenas, I., Crespo, P. & Célleri, R. (2012). Efectos Sobre las Propiedades Físicas y Químicas de los Suelos por el Cambio de la Cobertura Vegetal y Uso del Suelo: Páramo de Quimsacocha al Sur del Ecuador. *Suelos Ecuatoriales*, 42 (2): 138-153.

Quirk, J. (1994). Interparticle forces: A basis for the interpretation of soil physical behavior. In vol. 53 *Advances in Agronomy*, pp. 121-183. Orlando: Academic Press.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Rawls, W. J., Pachepsky, Y. & Shen, M. H. (2001). Testing soil water retention estimation with the MUUF pedotransfer model using data from the southern United States. *Journal of Hydrology*, 251 (3): 177-185. doi: https://doi.org/10.1016/S0022-1694(01)00467-X.

Rawls, W. J., Pachepsky, Y. A., Ritchie, J. C., Sobecki, T. M. & Bloodworth, H. (2003). Effect of soil organic carbon on soil water retention. *Geoderma*, 116 (1): 61-76. doi: https://doi.org/10.1016/S0016-7061(03)00094-6.

Rawls, W. J., Nemes, A. & Pachepsky, Y. (2004). Effect of soil organic carbon on soil hydraulic properties. In Development of Pedotransfer Functions in Soil Hydrology, vol. Volume 30 *Developments in Soil Science*, pp. 95-114: Elsevier.

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. *Chemical Geology*, 429 (Supplement C): 75-84. doi: https://doi.org/10.1016/j.chemgeo.2016.03.010.

Riley, H. (1979). Sammenhengen mellom jordas vannholdende evne og dens mekaniske sammenheng, moldinnhold og volumvekt. *Forskning og forsøk i landbruket*.

Romano, N. & Chirico, G. B. (2004). The role of terrain analysis in using and developing pedotransfer functions. In Development of Pedotransfer Functions in Soil Hydrology, vol. Volume 30 *Developments in Soil Science*, pp. 273-294: Elsevier.

Salter, P. J. & Williams, J. B. (1969). The influence of texture on the moisture characteristics of soil. *Journal of Soil Science*, 20 (1): 126-131. doi: doi:10.1111/j.1365-2389.1969.tb01561.x.

Sequeira, C. H., Wills, S. A., Seybold, C. A. & West, L. T. (2014). Predicting soil bulk density for incomplete databases. *Geoderma*, 213: 64-73. doi: https://doi.org/10.1016/j.geoderma.2013.07.013.

Shoji, S., Nanzyo, M., A. Dahlgren, R. & Quantin, P. (1996). Evaluation and proposed revisions of criteria for Andosols in the World Reference Base for Soil Resources. *Soil Science*, 161. doi: 10.1097/00010694-199609000-00005.

Tapia, F. M. & Pacheco, C. G. (2015). *Efectos de las plantaciones de pino (pinus patula) en las propiedades hidrofísicas de los horizontes ándicos de los suelos de páramo en dos zonas de la cuenca alta del Río Paute*: University of Cuenca.

Tate, K. R. & Theng, B. K. G. (1980). Organic matter and its interactions with inorganic soil constituents. In Theng, B. K. G. (ed.) *Soils with variable charge*, pp. 225-249. Lower Hutt, New Zealand: Soil Bureau.

Tietje, O. & Tapkenhinrichs, M. (1993). Evaluation of pedo-transfer functions. *Soil Science Society of America Journal*, 57 (4): 1088-1095.

Tokashiki, Y. & Wada, K. (1975). Weathering implications of the mineralogy of clay fractions of two Ando soils, Kyushu. *Geoderma*, 14 (1): 47-62. doi: https://doi.org/10.1016/0016-7061(75)90012-9.

Tomasella, J., Pachepsky, Y., Crestana, S. & Rawls, W. J. (2003). Comparison of Two Techniques to Develop Pedotransfer Functions for Water Retention. *Soil Science Society of America Journal*, 67 (4): 1085-1092. doi: 10.2136/sssaj2003.1085.

van Beers, W. F. J. (1970). *The Auger-hole method: a field measurement of the hydraulic conductivity of soil below the water table*. Wageningen, The Netherlands: International Institute for Land Reclamation and Improvement (ILRI).

van Dam, J. C., Stricker, J. N. M. & Droogers, P. (1994). Inverse Method to Determine Soil Hydraulic Functions from Multistep Outflow Experiments. *Soil Science Society of America Journal*, 58 (3): 647-652. doi: 10.2136/sssaj1994.03615995005800030002x.

van Genuchten, M. (1980). A Closed-form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils. *Soil Science Society of America*, 44: 892-898. doi: 10.2136/sssaj1980.03615995004400050002x.

Van Genuchten, M., J. Leij, F., R. Yates, S. & R. Williams, J. (1991). *The RETC Code for Quantifying Hydraulic Functions of Unsaturated Soils*, vol. 83.

van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., Montzka, C., Nemes, A., Pachepsky, Y. A., Padarian, J., et al. (2017). Pedotransfer Functions in Earth System Science: Challenges and Perspectives. *Reviews of Geophysics*, 55 (4): 1199-1256. doi: doi:10.1002/2017RG000581.

Veihmeyer, F. & Hendrickson, A. (1927). The relation of soil moisture to cultivation and plant growth. *Proc. 1st Intern. Congr. Soil Sci*, 3: 498-513.

Vereecken, H., Weynants, M., Javaux, M., Pachepsky, Y., Schaap, M. G. & Genuchten, M. T. v. (2010). Using Pedotransfer Functions to Estimate the van Genuchten–Mualem Soil Hydraulic Properties: A Review. *Vadose Zone Journal*, 9 (4): 795-820. doi: 10.2136/vzj2010.0045.

von Post, L. & Granlund, E. (1926). *Södra Sveriges torvtillgångar*. Sveriges Geologiska Undersökning. Ser. C 335.

Wada, K. (1977). Active Aluminum in Kuroboku Soils and Nonand Para-Crystalline Clay Minerals. *Journal of the Clay Science Society of Japan (in Japanese)*, 17 (4): 143-151. doi: 10.11362/jcssjnendokagaku1961.17.143.

Wösten, J., Pachepsky, Y. A. & Rawls, W. (2001). Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. *Journal of hydrology*, 251 (3-4): 123-150.

Wösten, J. H. M., Finke, P. A. & Jansen, M. J. W. (1995). Comparison of class and continuous pedotransfer functions to generate soil hydraulic characteristics. *Geoderma*, 66 (3): 227-237. doi: https://doi.org/10.1016/0016-7061(94)00079-P.

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway