Pedotransfer Functions: Improving Predictions through Machine Learning and Nonlinear Least Squares Approaches Coupled with Quantile Regression

by

Adrienne Arbor

BSc (Geography), Simon Fraser University, 2020

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Geography
Faculty of Environment

© Adrienne Arbor 2023
SIMON FRASER UNIVERSITY
Fall 2023

Copyright in this work is held by the author. Please ensure that any reproduction or re-use is done in accordance with the relevant national copyright legislation.

Declaration of Committee

Name:	**Adrienne Arbor**
Degree:	**Master of Science (Geography)**
Title:	**Pedotransfer Functions: Improving Predictions through Machine Learning and Nonlinear Least Squares Approaches Coupled with Quantile Regression**
Committee:	**Chair: Kirsten Zickfeld** Professor, Geography
	Margaret Schmidt Co-Supervisor Associate Professor, Geography
	Brandon Heung Co-Supervisor Adjunct Faculty, Geography
	Bing Lu Committee Member Assistant Professor, Geography
	Sean Smukler Committee Member Associate Professor, Land and Food Systems University of British Columbia
	Maja Krzic Examiner Associate Professor, Land and Food Systems University of British Columbia

Abstract

Digital soil mapping requires input data, which often are sourced from legacy soil datasets. These datasets may be incomplete and require the use of pedotransfer functions (PTFs) to estimate the missing soil attribute values. Two methods of increasing the accuracy of PTFs are explored: the use of nonlinear least squares (NLS) to recalibrate existing equation-based functions; and the machine learner Random Forest (RF) to develop new PTFs. The target attribute used as a case study was bulk density (BD), which is a soil variable often missing in legacy soil datasets. To test the effectiveness of the NLS method in recalibrating existing PTFs, 73 PTFs from literature were tested on three regional datasets, two from British Columbia (BC) and one from Ontario. Improvement in accuracy was gauged through the comparison of root mean square error (RMSE) and concordance correlation coefficient (CCC) values determined before and after recalibration. Results showed that the accuracy of almost every PTF improved; PTFs with fewer variables and those recalibrated on the largest dataset showed the highest accuracy. The machine learner RF was also used to develop PTFs. Eleven variables were available in the legacy dataset from BC used as a case study region, and all possible combinations of these variables were used to create 512 models for predicting BD. After testing the models, they were ranked based on their CCC value, and showed a range of 0.92 for the best performing model, to 0.51 for the lowest ranked model. The number of horizons which could be estimated by each model also varied, as many of the variables were limited in their availability. To estimate missing BD values in the dataset, models were chosen on their performance and number of horizons which could be estimated, with 27 models used to estimate the missing BD values. Lastly, as most developed PTFs lack accompanying uncertainty estimates, quantile regression (QR) was used to address this gap. PTF uncertainty was shown to be related to the size of the training dataset used as well as the input variables. A framework that coupled a quantile regression approach both with PTF recalibration and with PTF development was constructed that produced region specific PTFs along with uncertainty estimates; the predictions were used to fill legacy soil datasets.

Keywords: Pedotransfer Function; Machine Learning; Nonlinear Least Squares; Quantile Regression; Legacy soil data; Random Forest

This thesis is dedicated to my family.

My parents, Heather and Dave, for taking me in, brushing me off, and helping me stand again. You encouraged me to dream big, and to believe in myself. You picked up Jasper from school, you put up with pets, seasonal decorations, and my tendency to clutter every horizontal surface. I am so lucky to have you both.

My irreverent partner in adventure, Sean. You listened patiently to my problems, were understanding when I brought my laptop on every vacation, and you inspired me with your resilience and persistence. There is no one I would rather have by my side – from the campfire to the kitchen to the next corner of BC we explore.

My son, Jasper – you are the spark, and the bringer of treasure to my life. I admire your thoughtfulness and determination, and I am so proud of you. You have been so patient with the ups and downs of having a mom in school, and make me laugh every day. You are wise beyond your years, kiddo.

I love you all, and I am so grateful for your support throughout this journey.

Acknowledgements

I am so grateful for the support of my supervisor, Dr. Margaret Schmidt. She took a chance on me as an undergraduate, and hired me as a work study student in the soil lab. She encouraged me to apply for awards despite my GPA at the time, and wrote letters of recommendation. She provided many opportunities, trusted me, and supported me on every step of this journey. I can't thank her enough, because I truly would not have had the chance to write this thesis without her.

I would also like to thank Dr. Brandon Heung for his guidance. I always felt like I had someone to turn to for technical questions, and he pushed me to be a better writer and to think critically about what I was doing. I am grateful for his support, the time he took to coach me through scholarship applications, and his patience.

I am also thankful to Jin Zhang, who wrote code, answered questions, and encouraged me along the way. Her work was integral to this thesis, and we had many enjoyable conversations over the past several years.

I would also like to thank Babak Kasraei, with whom I worked in the soil lab, and as a teaching assistant. Babak took the time to show me his research, and was a patient supervisor.

As well, I want to extend thanks to the working group members, especially Chuck Bulmer and Deepa Filatow, for their insight and advice through this process.

Also, those members of the geography department who make everything happen behind the scenes – Anke Baker, Curtis Platson, and Joyce Chen. Joyce was always patient with my requests for keys or questions about how to use the copier, and it was treat to say hi to Theo in the office. Curtis was ever helpful and patient, and guided me through the confusions of paperwork. And Anke was continuously thoughtful with her TA assignments, which was a kindness for which I am grateful.

Lastly, to all the professors and instructors throughout my SFU journey, who provided support, inspiration, and made me feel at home in Geography.

Table of Contents

List of Tables

List of Figures

List of Acronyms

BD	British Columbia
BD	Bulk Density
CCC	Concordance Correlation Coefficient
CF	Coarse Fragment
DSM	DSM
ML	Machine Learning
MPI	Mean Prediction Interval
NLS	Nonlinear Least Squares
OC	Organic Carbon
OM	Organic Matter
PICP	Prediction Interval Confidence Probability
PM	Parent Material
PTF	Pedotransfer Function
QR	Quantile Regression
RF	Random Forest
RMSE	Root Mean Square Error
RMSPE	Root Mean Square Prediction Error
SDPE	Standard Deviation of the Predicted Error

Chapter 1.

Introduction

Soil data are needed for many applications. It may be used to produce maps of soil attributes or used as input for modelling processes. The products which soil data are used for, such as maps, are themselves used as decision-making tools; to convey information and discover soil processes. Soil data come from a variety of sources, ranging from small, regional case studies, to large national datasets; and data are constantly being produced. Globally, most countries have generated soil data, through soil surveying and mapping (Arrouays et al., 2017). Through multiple initiatives, there has been a push to incorporate existing soil data into harmonized, global datasets, such as the GlobalSoilMap project (Arrouays, 2017). The data can be used in digital soil mapping projects, where the use of predictive modeling and uncertainty assessments have been identified as current topics (Arrouays, 2020).

Lagacherie and McBratney (2007) defined digital soil mapping (DSM) as “the creation, and population of spatial soil information systems by the use of field and laboratory observational methods, coupled with spatial and non-spatial soil inference systems”. DSM began in the 1990s (Minasny and McBratney, 2016) and has expanded with the development of other technologies, such as geographic positioning systems and remote sensing (Brevik et al., 2016). DSM projects require input in the form of data; good data are essential in the creation of an accurate DSM, but acquiring good data is often a limiting factor (Lagacherie, 2008). Data may come from legacy sources, such as soil maps or existing soil databases, and it may also be acquired from environmental observations. Legacy data are often incomplete, however; the data may come from different studies, with measurements made using different methods, and collected over long time periods and for different purposes. Further, data may not be harmonised, and spatial locations and soil descriptions may be imprecise (Lagacherie, 2008).

To utilize existing data to its fullest extent, it is useful to “fill the gaps” in the dataset using pedotransfer functions (PTFs), which are methods of predicting soil attributes, and which describe relationships between soil variables. When soil data are missing, a PTF can be applied to estimate the missing values. In the words of the soil

scientist who coined the term, they “translate the data we *have* to the data we *need*” (Bouma, 1989). McBratney et al. (2002) defined PTFs as “predictive functions of certain soil properties from other easily, routinely, or cheaply measured properties”. McBratney et al. (2003) noted that relationships between soil attributes, and only soil attributes, without the inclusion of spatial position, fall under the definition of pedotransfer functions (PTFs). What might be considered PTFs, but which include variables categorized as one of the soil formation factors such as parent material and which have a spatial component, would be defined as soil spatial prediction functions (SSPFs), rather than as PTFs. The authors contend that a PTF, such as s = f(r), where a soil attribute is predicted based on topography, “extends the definition too far”. However, with machine learning becoming prevalent in PTF development, more and more PTFs incorporate soil formation factors as variables.

There are many PTFs available in the literature. While much of PTF development has focused on soil hydraulic properties and bulk density, there are also PTFs available for soil organic carbon (SOC; Benke et al., 2020; Mwango et al., 2019; Fernández-Ugalde and Tóth, 2017; Zinn et al., 2005), for sand, silt, and clay content (Levi et al., 2017), for sand or silt percent (Furze and Arp, 2018). PTFs have been developed for many regions around the globe, for varying types of soil and environmental conditions. However, it is not recommended to use a PTF outside the region or soil type for which it was developed (McBratney et al., 2002). When faced with a choice of choosing a literature PTF or developing a new PTF, a third option exists: recalibrating an existing PTF to suit the study region. With recalibration, an existing equation is used, for which new coefficients are generated. This fits the equation better to the dataset under consideration. Recalibration can be accomplished through the use of non-linear least squares (NLS), an iterative process which provides an averaged best estimate of coefficients appropriate for the dataset.

The option of developing a new PTF can be achieved using machine learning. There are many types of machine learners available; examples of machine learners which have been used to estimate soil bulk density include boosted regression trees (Martin et al., 2009; Gharahi Ghehi et al., 2012); artificial neural networks (Al-Qinna and Jaber, 2013); generalized boosted regression (GBM), Chen et al., 2018; Jalabert et al., 2010; k-nearest neighbour (Gharahi Ghehi et al., 2012; Botula et al., 2015) and support vector machines (Guo et al., 2019). A machine learner which has shown good results in

many modeling applications (Boulesteix et al., 2012) is Random Forest (RF; Breiman, 2001) which is an ensemble of tree-based learners. Advantages of RF include its ability to resist overfitting (Breiman, 2001); it can handle high dimensional data, where the number of predictors is greater than the number of observations (Grimm et al., 2008); and it can handle continuous and categorical variables (Ließ et al., 2012), which is an advantage when including environmental variables.

The inclusion of uncertainty estimates when developing PTFs is important, both to have an uncertainty estimate with any predicted variable, and because the predictions produced by PTFs are often used in further modelling work. The uncertainty of one variable can depend on the uncertainty of another variable, and so propagate through the model (Heuvelink and Brown, 2007). For example, when modelling soil acidification, Finke et al. (1999) investigated the uncertainty associated with categorical data used as input to the model, and found that the error of the input data had a definite effect on the uncertainty of the model results (Finke et al., 1999). PTFs have not usually been accompanied by an uncertainty estimate (McBratney et al., 2002), but this has been recognized as an issue to be addressed by several studies (Tranter et al., 2010; Malone et al., 2011; Van Looy et al., 2017). There have been different approaches to quantifying uncertainty, in both DSM and PTF literature. For example, Goovaerts (2001) used two techniques, kriging-based and simulation-based, to estimate the uncertainty of continuous soil attributes. One available method is Quantile Regression (QR; Koenker and Bassett, 1978) which is especially useful for applications where the residuals may not be normally distributed, and it allows for exploration of the residuals beyond the mean; QR is also not sensitive to outliers (Koenker, 2017), and it is computationally efficient.

1.1. Background Information

This section provides background information for Chapters 2 and 3. Section 1.1.1 provides motivation for the research, in the importance of soil; 1.1.2 focuses on the development of soil science and soil surveying, which provided the basis and data which are used in the research; 1.1.3 describes issues with legacy soil data; section 1.1.4 describes the development of PTFs; 1.1.5 details the method used in Chapter 2 to recalibrate existing equation-based PTFs; Section 1.1.6 discusses using a machine learning approach to predicting soil attributes, as expanded on in Chapter 3; and lastly,

section 1.1.7 looks at the importance of including uncertainty estimates when generating pedotransfer functions, and is further demonstrated in both Chapter 2 and Chapter 3.

1.1.1. Soil and Humanity

As agriculture became important to humanity, so did soil. Agricultural tools used 11,000 years ago have been found in Iraq (Brevik and Hartemink, 2010). The same region has yielded evidence of irrigation from 9,500 BP, and an early type of plough called the ard, from 6000 to 4000 years BP (Brevik and Hartemink, 2010). In China, rice grains dating to 9100 years BP were found in Pengtoushan (Gong et al., 2003); millet and rice were being cultivated 7000 to 6000 years BP (Gong et al., 2003). Evidence of ancient agriculture has also been found dating to 7,500 years BP in Poland (Brevik and Hartemink, 2010); in Uzbekistan 6,000 years BP (Brevik and Hartemink, 2010); and in India 4-5,000 years BP (Miller and Schaetzl, 2014). While our ancestors may not have understood the chemical and physical qualities of the soil in the same way we do today, they did comprehend the importance of soil to their survival. For example, ancient peoples often were the cause of soil erosion through agricultural practices (Dotterweich, 2013); however, there is also evidence from around the globe that solutions to this problem were implemented – people from the Phoenicians to the Maya and Inca built terraces to prevent erosion (Brevik and Hartemink, 2010).

So too now, we face the problem of degraded soil; for example, soil carbon has been lost through erosion, decomposition, and leaching caused by land use change and land management practices (Lal, 2018). Loss of carbon from the soil is one source of atmospheric CO_2; in 2021, the total CO_2 emissions from fossil fuels was 9.9 +/- 0.5 Gt C/year, an increase of 0.46 Gt C/year from the previous year; and the total CO_2 emissions from land use change was 1.1 +/- 0.7 Gt C/year, with both emissions estimations including the cement carbonation sink (Friedlingstein et al., 2022). The increased CO_2 in the atmosphere is a major contributor to global climate change (Lal, 2018).

The idea that carbon influences the global climate is not a new one. Although he was interested in ice ages rather than global warming, in 1896 Svante Arrhenius calculated that if the carbon dioxide concentration of Earth's atmosphere were to double, the global temperature would increase by 4.0 – 6.1°C (Schils, 2011). In 1750, the

concentration of carbon dioxide (CO_2) in the atmosphere was estimated to be 278 ppm, while in 2021 it was 414.7 +/- 0.1 ppm (Friedlingstein et al., 2022). Soil is the largest terrestrial store of carbon, holding 1550 Gt of organic carbon and 950 Gt of inorganic carbon (Lal, 2004a). Through management, there is the potential to decrease atmospheric CO_2 by increasing carbon stored in soil (Lal, 2004a), to mitigate the 31% increase of CO_2 concentration in the atmosphere since 1750 (Lal, 2004b).

1.1.2. Soil Surveys and Science

In China, land quality data and associated crop type was recorded from 300 AD (Miller and Schaetzl, 2014). Chernozem soils were mapped by the Russian Dokuchaev during the late 1800s, and the US Soil Survey began mapping soils in 1899 (Miller and Schaetzl, 2014). The first Canadian soil survey report was published in 1923, based on soil surveying done from 1914 to 1920 in southwestern Ontario (McKeague and Stobbe, 1978). In BC, soil surveying first occurred in 1926, and was focused on determining whether the land was arable or suitable for inclusion in forest reserves (McKeague and Stobbe, 1978). Soil surveying was carried out by the provinces, and as a result there were differences in the scale of surveys and mapping units used (McKeague and Stobbe, 1978). To develop a cohesive national survey system, in 1945 the National Soil Survey Committee was formed (McKeague and Stobbe, 1978). A five-category classification system based on soil associations was adopted (Anderson and Smith, 2011). In 1955 a classification system based on soil taxonomy was initiated and accepted for use in 1960 (Anderson and Smith, 2011). This system was formalized as the System for Soil Classification in Canada, published in 1974 (Anderson and Smith, 2011).

While soil science has long focused on agricultural applications, in more recent years there has been broadening of the scope of soil science, with environmental issues receiving more attention (Hartemink and McBratney, 2008). Soil has many important functions, from providing a medium for plants to grow in, holding and filtering water, as a component of construction, to regulating the carbon cycle, among many others (Omuto et al., 2013). To study these functions, data are required; and there have been multiple initiatives undertaken to centralize, harmonize, and make soil data accessible (Omuto et al., 2013). However, legacy data are often inconsistent, with varying amounts of accompanying observations available for any given data point.

1.1.3. Legacy data

Legacy data may be available from soil maps or from soil profile data gathered from sampling (Lagacherie, 2008). When in the form of soil maps, this information must be disaggregated. In the case of soil profile data, there can be issues such as lack of harmonization, missing data, and imprecise location information (Lagacherie, 2008); data may have been collected for different objectives and through different methodologies (Krol, 2008) One issue is the location of sampling sites, which were chosen for specific purposes such as investigating agricultural potential; this can result in sampling bias in the data (Carré et al., 2007). However, there may not be an option to collect new data, and so legacy data should be used to the greatest extent possible (Lagacherie, 2008).

There are numerous studies which utilize legacy data. Bui et al. (2006) used legacy data from a national Australian soil database to model multiple soil attributes using piecewise linear decision trees. In Saskatchewan, Canada, Sorenson et al. (2021) used legacy soil data in conjunction with remotely sensed imagery from the Landsat 5 satellite to model soil organic carbon, clay, and cation exchange capacity. Many studies have drawn from large, national or international soil databases; Sequeira et al. (2014) used the USDA-NRCS National Soil Survey Center database to compare approaches to predicting bulk density; Vaysse and Lagacherie (2015) used legacy soil data in France, including a map of soil classes and a dataset of soil profiles, as well as another dataset of previously collected composite profiles which was used for validation. For that study, they used four different models to predict 8 soil attributes: clay, silt, sand, coarse fragment, organic carbon, pH, CEC, and depth to bedrock. When legacy datasets are incomplete, the missing values may be estimated through PTFs to provide a complete dataset for further digital soil mapping endeavours. This was the approach taken by Silatsa et al. (2020) when comparing different methods of predicting soil carbon stocks in Cameroon. Bulk density values missing in the dataset were estimated using Minasny and Hartemink's (2011) method of first predicting the bulk density of the mineral soil, then inputting that value into the PTF developed by Adams (1973).

1.1.4. Pedotransfer Function Development

The term “pedotransfer function” was created by Bouma (1989) and based on the term “pedofunctions” used by Lamp and Kneib in 1981 (Bouma, 1989), and “transfer functions” from Bouma and van Lanen in 1987 (Bouma, 1989). As defined, “Pedotransfer functions relate different soil characteristics and properties with one another or to land qualities.” (Bouma, 1989). Two types of PTFs were identified: class and continuous. Continuous functions are equations which use continuous variables such as percent sand as input variables; class functions relate a soil attribute to a class, such as soil taxonomic horizon.

While the term was novel, the idea was not. It can be argued that the earliest PTF was van Bemmelen’s conversion factor, published in 1889, which relates soil organic matter to soil organic carbon, using a factor of 1.724 (Minasny et al., 2020). However, van Bemmelen referred to the conversion factor as “the factor of Wolff: 1.724” (Minasny et al., 2020); Emil Theodor von Wolff likely used this factor based on the work of Carl Sprengel, who published works in 1826 and 1827 which stated that humic acid was composed of 58% carbon (Minasny et al., 2020). Although multiple studies have shown that the ratio of OC to OM is variable and this conversion factor overestimates OC, it is still used as the default today (Pribyl, 2010).

Later, regression analysis was used to identify relationships between soil variables and bulk density, such as by Eschner et al. (1957), who developed two equations for bulk density based on soil organic matter, for different depth intervals; as well as Curtis and Post (1964), Saini (1966), Jeffrey (1970), Stewart et al. (1970), Drew (1973), and Adams (1973), who all used organic matter or organic carbon to determine bulk density. As PTF development continued for different regions, other variables began to be included more frequently, although organic matter/organic carbon continued to be one of the most frequent variables included. PTFs continue to be developed using MLR, such as by Foldal et al. (2020) whose bulk density estimates had an RMSE of 0.190 g/cm^3; some studies had very good results, such as Obidike-Ugwu et al. (2022) whose bulk density PTF had an RMSE of 0.07 g/cm^3. An advantage of using MLR is its ease of use (Obidike-Ugwu et al., 2022). Whichever approach is used to develop a PTF, a PTF should “not predict something that is easier to measure than the predictor” (McBratney et al., 2002).

1.1.5. Recalibrating Existing Equation-Based PTFs

While many PTFs have been developed for different regions and soil conditions, there has been general consensus that a PTF should not be applied outside of the region or conditions for which it was developed (McBratney et al., 2002; De Vos et al., 2005; Benites et al., 2007; Casanova et al., 2016). An alternative to developing a new PTF or to using an existing PTF is recalibration, where for an existing equation, the coefficients of the variables can be adjusted to better suit the data in the study. A method that has previously been used to recalibrate existing PTFs is nonlinear least squares (De Vos et al., 2005; Nanko et al., 2014; Chen et al., 2018). Khodaverdiloo et al. (2022) used nonlinear least squares, and compared the results to unrecalibrated PTFs, as well as to PTFs developed using MLR. They tested these on their whole dataset (n=360), as well as smaller subsets of the dataset. They noted that the size of the dataset affects the results; as the dataset size increased, accuracy decreased.

Nonlinear least squares (NLS) is a method of fitting a nonlinear function to a set of data points, which minimizes the sum of the squares of the residual (Sun and Yuan, 2006). Like linear least squares, NLS fits an equation to a dataset by finding the optimal parameters (Johnson and Frasier, 1985). Johnson (2008) identifies the difference between a linear and nonlinear least-squares fit is “if the second and higher order derivatives of the fitting function with respect to the parameters being estimated are all equal to zero, then the fit is a linear fit. If any of these derivatives are not equal to zero then it is a nonlinear fit.” There are multiple weighted NLS fitting algorithms (Johnson, 2008), including the Levenberg-Marquardt method, the Gauss-Newton approach, and the Quasi-Newton method (Sun and Yuan, 2006). The methods are iterative and act in a stepwise way to reach the optimal values for the parameters of the equation; this is known as convergence (Ritz and Streibig, 2008). When applying NLS, a model must be specified (Bates and Watts, 1988); an existing literature PTF equation can be used. It is important to provide the best starting values for the parameters, as this will result in convergence being reached more quickly (Bates and Watts, 1988). Convergence may not always be reached, and this could be due to a number of factors: there may be too many parameters in the model, there could be too little data in some areas of the function, the starting values may be poor or have the wrong sign (Bates and Watts, 1988).

1.1.6. Machine Learners and Random Forest

Wadoux et al. (2020) defined machine learning as “the computer-assisted practice of using data-driven (and mostly non-linear) statistical models which resort to a large amount of input data to learn a pattern and make a prediction.” Machine learning was identified as a trend in digital soil mapping (Arrouays et al., 2020) and has also been increasingly used in PTF development. An advantage of using machine learning to predict a soil attribute is that there is no assumption of the character of the relationship between the target variable and the input variables (Wadoux et al., 2020). However, unlike regression, machine learners do not produce a readily understandable equation (Sequeira et al., 2014). Instead, the models they produce have been referred to as “black boxes” (Breiman, 2001), and are difficult to interpret (Wadoux et al., 2020).

With the availability of machine learners, many studies have compared multiple approaches to PTFs – often multiple linear regression (MLR) and one or several machine learners. For example, Katuwal et al. (2020) found that all of the machine learners tested - RF, regression rules, and ANN - outperformed MLR in predicting bulk density; Schillaci et al. (2021) compared stepwise MLR, backward stepwise MLR, and ANN, and also found that ANN performed better than either of the MLR approaches. Other studies have compared different machine learners to each other to estimate bulk density; these include Gunarathna et al. (2019) who compared ANN, kNN, and RF; and Gharahi Ghehi et al. (2012) whose study compared kNN and BRT. Of multiple machine learners assessed, including RF, ANN, and support vector machine, Zihao et al. (2022) found that RF performed the best, with the model’s prediction accuracy having an RMSE of 0.147 g/cm^3.

RF has been used by many studies to estimate soil bulk density; for example, Ramcharan et al. (2017) used RF to estimate bulk density, with a resulting RMSPE of 0.13 g/cm^3 for the estimates produced. Other studies using RF include Hikouei et al. (2021); Akpa et al. (2016); de Souza et al. (2016); Palladino et al. (2022). RF (Breiman, 2001) is a type of decision tree algorithm, where multiple trees are grown. Trees are grown on bootstrapped data, and the results from the trees are aggregated. For classification, the majority vote of the trees is the result; in regression, the predictions of the trees are averaged (Hastie et al., 2009). Certain characteristics of RF make it an attractive machine learner to use: it does not overfit (Breiman, 2001); it is nonparametric;

it can be used for problems which have more variables than observations; it can handle categorical and continuous variables; and it has few parameters to configure (Genuer and Poggi, 2020).

1.1.7. Uncertainty and Quantile Regression

Wadoux et al. (2020) reported that of 150 studies which used machine learning for digital soil mapping purposes, approximately 30% included uncertainty quantification with their predictions. The authors recommend that uncertainty originating from both the data and the model be reported in future DSM studies (Wadoux et al., 2020). Model error has multiple sources; Jansen (1998) identified input uncertainty, uncertainty associated with the model structure, and system randomness. Prediction error may be underestimated, because the focus of uncertainty analysis is on input uncertainty, and other sources of uncertainty may be neglected (Jansen, 1998). For PTFs specifically, McBratney et al. (2002) pointed out the lack of uncertainty estimates and identified it as a problem to be addressed.

The probability distribution of error is often assumed to be a certain shape, such as a normal or Poisson distribution (Heuvelink and Brown, 2007). These characteristic distributions need few parameters to describe them – the normal distribution requires a mean and standard deviation, while the Poisson distribution requires the mean (Heuvelink and Brown, 2007). But the assumption that the probability distribution function follows a characteristic distribution may not be true; it may be non-parametric, and so require a different method of quantifying and conveying the uncertainty (Heuvelink and Brown, 2007). An advantage of QR is that it does not assume that the error distribution has any parametric form (Cade and Noon, 2003).

While QR has been used in other fields, such as flood forecasting (Amina and Chithra, 2023) and medical research (Beyerlein, 2014), its application in soil science has been limited. Lombardo et al. (2018) used QR when modeling soil organic carbon; van Zijl et al. (2014) used QR to investigate the relationships between soil properties and soil dispersion; Kasraei et al. (2021) coupled QR with multiple machine learners to assess model performance and uncertainty estimates.

1.2. Research Problem

The goal of the research in both chapters was to investigate methods of improving the accuracy of PTFs. These PTFs are necessary to fill the gaps in legacy soil datasets, so that existing data can be used to the greatest extent possible in further digital soil mapping endeavours. There are PTFs available for many soil attributes, including pH; remaining phosphorus, which is used as an anion-adsorption index (Cagliari et al., 2011); cation exchange capacity (Liao et al., 2015; Krogh et al., 2000); electrical conductivity (Benke et al., 2020), and more. As a case study variable, bulk density was chosen, as it had low coverage in available datasets, is required for soil carbon stock estimations and other calculations, such as soil water matric potential (Box and Taylor, 1962), and has many existing PTFs with which results can be compared. Two methods, recalibration of existing equation based PTFs using NLS, and development of new PTFs using machine learning, were investigated as means of improving PTF performance for legacy dataset gap filling. For both methods, QR was applied as a method of generating uncertainty estimates, as many PTF studies lack uncertainty estimates.

1.3. Research Objectives

The research objectives were as follows:

1) Determine the potential for recalibrating existing equation based PTFs to improve accuracy using NLS on regional datasets

2) Generate new PTFs through a machine learning approach using RF

3) Produce uncertainty estimations for both recalibrated and new PTFs through a quantile regression approach

1.4. Thesis Overview

The thesis has four chapters: Chapter 1 is the introduction, which positions the research in context and provides background information on the material presented in Chapters 2 and 3.

Chapter 2 addresses multiple research objectives: 1) identifying existing equation-based PTFs; 2) recalibrating these PTFs through the use of nonlinear least squares; 4) producing uncertainty estimates for each recalibrated PTF. Existing equation-based PTFs were identified in the literature, and selected for having the case study variable, bulk density, as their target variable, as well as using input variables which were available in the case study datasets. Datasets from two regions of Canada, the province of BC, and an area of southwestern Ontario, were used to test the existing PTFs in their literature form, and then compare these results to the accuracy of the PTFs after recalibration. Recalibration was conducted through nonlinear least squares (NLS), which produces new existing coefficients for existing model forms. Coupled with recalibration, uncertainty estimates were generated through a quantile regression (QR) approach. The validated results of the recalibration, expressed through root mean square error (RMSE) and the concordance correlation coefficient (CCC), as well as the uncertainty estimates as conveyed through PICP and MPI graphs, were used to identify the PTFs with the highest accuracy and lowest uncertainty. It was found that when recalibrated on larger datasets, PTFs showed lower uncertainty values. The PTFs which had the highest accuracy when recalibrated included those with fewer variables, and minimal transformations to those variables; further, organic carbon (OC) was found to be the most important variable for bulk density prediction.

Chapter 3 address the research objectives 3) generating a new PTF through a machine learning approach; 4) producing uncertainty estimates for new PTFs; and 6) developing a method for gap filling legacy soil datasets. To generate new PTFs, all possible combinations of the variables available in the case study dataset were determined, resulting in over 500 possible model forms. Using the machine learner Random Forest, these model forms were then tested on the case study dataset, which covered the province of BC. Testing the models for accuracy was coupled with quantile regression to produce uncertainty estimates for each model. The models were then ranked for accuracy, based on the CCC value; these models were then applied to the dataset to estimate missing bulk density values.

Chapter 4 is the conclusion. Conclusions drawn from the research are presented: the results from Chapter 2, where NLS was used to recalibrate existing PTFs, showed that for almost every PTF tested, an NLS approach increased the accuracy of the PTF; the results from Chapter 3, where RF was used to produce new PTFs using a variety of

input variables, showed that this approach resulted in PTFs with high accuracy. Further, it was shown that QR can quantify and communicate the uncertainty of the PTFs resulting from both approaches. Limitations with the research were also addressed, such as the limited amount of data available to train new models; the effect of the accuracy of the measured value of the target variable on the model results; the availability and quality of input variables; the importance of different input variables to the production of more accurate models; and potential ways of improving PTFs, such as the incorporation of environmental variables, and testing multiple machine learners to determine which one produces the most accurate PTF.

1.5. References

Adams, W.A. 1973. The effect of organic matter on the bulk and true densities of some uncultivated podzolic soils. Journal of Soil Science, **24(1):** 10-17.

Akpa, S.I.C., Ugbaje, S.U., Bishop, T.F.A., and Odeh, I.O.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:** 644-658.

Al-Shammary, A.A.G., Kouzani, A.Z., Kaynak, A., Khoo, S.Y., Norton, M., and Gates, W. 2018. Soil bulk density estimation methods: A review. Pedosphere, **28(4):** 581-596.

Al-Qinna, M.I., and Jaber, S.M. 2013. Predicting soil bulk density using advanced pedotransfer functions in an arid environment. Transactions of the ASABE, **56(3):** 963-976.

Amina, M.K., and Chithra, N.R. 2023. Predictive uncertainty assessment in flood forecasting using quantile regression. H_2Open Journal, **6(3):** 477-492.

Anderson, D.W., and Smith, C.A.S. 2011. A history of soil classification and soil survey in Canada: Personal perspectives. Canadian Journal of Soil Science, **91(5):** 675-694.

Arrouays, D., Leenaars, J.G.B., Richer-de-Forges, A.C., Adhikari, K., Ballabio, C., Greve, M., Grundy, M., Guerrero, E., Hempel, J., Hengl, T., Heuvelink, G., Batjes, N., Carvalho, E., Hartemink, A., Hewitt, A., Hong, S.-Y., Krasilnikov, P., Lagacherie, P., Lelyk, G., Libohova, Z., Lilly, A., McBratney, A., McKenzie, N., Vasques, G.M., Mulder, V.L., Minasny, B., Montanarella, L., Odeh, I., Padarian, J., Poggio, L., Roudier, P., Saby, N., Savin, I., Searle, R., Solbovoy, V., Thompson, J., Smith, S., Sulaeman, Y., Vintila, R., Rossel, R.V., Wilson, P., Zhang, G.-L., Swerts, M., Oorts, K., Karklins, A., Feng, L., Ibelles Navarro, A.R., Levin, A., Laktionova, T., Dell'Acqua, M., Suvannang, N., Ruam, W., Prasad, J., Patil, N., Husnjak, S., Pásztor, L., Okx, J., Hallett, S., Keay, C., Farewell, T., Lilja, H., Juilleret, J., Marx, S., Takata, Y., Kazuyuki, Y., Mansuy, N., Panagos, P., Van Liedekerke, M., Skalsky, R., Sobocka, J., Kobza, J., Eftekhari, K., Kacem Alavipanah, S., Moussadek, R., Badraoui, M., Da Silva, M., Paterson, G., da Conceição Gonçalves, M., Theocharopoulos, S., Yemefack, M., Tedou, S., Vrscaj, B., Grob, U., Kozák, J., Boruvka, L., Dobos, E., Taboada, M., Moretti, L., Rodriguez, D. 2017. Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. GeoResJ, **14:** 1-19.

Arrouays, D., Poggio, L., Salazar Guerrero, O.A., Mulder, V.L. 2020. Digital soil mapping and GlobalSoilMap. Main advances and ways forward. Geoderma Regional, **21:** e00265.

Bates, D.M., and Watts, D.G. 1988. Nonlinear Regression Analysis and Its Applications. New York: Wiley.

Bayerlein, A. 2014. Quantile regression – opportunities and challenges from a user's perspective. American Journal of Epidemiology, **180(3):** 330-331.

Benites, V.M., Machado, P.L.O.A., Fidalgo, E.C.C., Coelho, M.R. and Madari, B.E. 2007. Pedotransfer functions for estimating soil bulk density from existing soil survey reports in Brazil. Geoderma, **139:** 90-97.

Benke, K.K., Norng, S., Robinson, N.J., Chia, K., Rees, D.B., Hopley, J. 2020. Development of pedotransfer functions by machine learning for prediction of soil electrical conductivity and organic carbon content. Geoderma, **366:** 114210.

Botula, Y.-D., Nemes, A., Van Ranst, E., Mafuka, P., De Pue, J., Cornelis, W.M. 2015. Hierarchical pedotransfer functions to predict bulk density of highly weathered soils in Central Africa. Soil Sci. Soc. Am. J., **79:** 476-486.

Boulesteix, A.-L., Janitza, S., Kruppa, J., and König, I.R. 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Technical Report Number 129, Department of Statistics, University of Munich.

Bouma, J. 1989. Using soil survey data for quantitative land evaluation. In: Stewart, B.A. (eds) Advances in Soil Science. Advances in Soil Science, vol. 9. Springer, New York, NY.

Box, J.E., Taylor, S.A. 1962. Influence of soil bulk density on matric potential. Soil Sci. Soc. Am. J., **26(2):** 119-122.

Breiman, L. 2001. Random Forests. Machine Learning, **45:** 5-32.

Brevik. E.C. and Hartemink, A.E. 2010. Early soil knowledge and the birth and development of soil science. Catena, **83:** 23-33.

Brevik, E.C., Calzolari, C., Miller, B.A., Pereira, P., Kabala, C., Baumgarten, A., Jordán, A. 2016. Soil mapping, classification, and pedologic modeling: History and future directions. Geoderma, **264:** 256-274.

Bui, E.N., Henderson, B.L., Viergever, K. 2006. Knowledge discovery from models of soil properties developed through data mining. Ecological Modelling, **191:** 431-446.

Cade, B.S. and Noon, B.R. 2003. A gentle introduction to quantile regression for ecologists. Front. Ecol. Environ., **1(8):** 412-420.

Cagliari, J., Veronez, M.R., Alves, M.E. 2011. Remaining phosphorus estimated by pedotransfer function. R. Bras. Ci. Solo, **35:** 203-212.

Carré, F., McBratney, A.B., Minasny, B. 2007. Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. Geoderma, **141:** 1-14.

Casanova, M., Tapia, E., Seguel, O., Salazar, O. 2016. Direct measurement and prediction of bulk density on alluvial soils of central Chile. Chilean Journal of Agricultural Research, **76(1):** 105-113.

Chen, S., Richer-de-Forges, A.C., Saby, N.P.A., Martin, M.P., Walter, C., and Arrouays, D. 2018. Building a pedotransfer function for soil bulk density on regional dataset and testing its validity over a larger area. Geoderma, **312:** 52-63.

Curtis, R.O., and Post, B.W. 1964. Estimating bulk density from organic-matter content in some Vermont forest soils. Soil Sci. Soc. Am. Proc., **28:** 285-286.

De Souza, E., Filho, E.I.F., Schaefer, C.E.G.R., Batjes, N.H., dos Santos, G.R., Pontes, L.M. 2016. Pedotransfer functions to estimate bulk density from soil properties and environmental covariates: Rio Doce basin. Scientia Agricola, **73(6):** 525-534.

De Vos, B., Van Meirvenne, M., Quataert, P., Deckers, J. and Muys, B. 2005. Predictive quality of pedotransfer functions for estimating bulk density of forest soils. Soil Sci. Soc. Am. J., **69:** 500-510.

Dotterweich, M. 2013. The history of human-induced soil erosion: geomorphic legacies, early descriptions and research, and the development of soil conservation – a global synopsis. Geomorphology, **201:** 1-34.

Drew, L.A. 1973. Bulk density estimation based on organic matter content of some Minnesota soils. Minnesota Forestry Research Notes, **243.**

Eschner, A.R., Jones, B.O., and Moyle, R.C. 1957. Physical properties of 134 soils in six northeastern states. Station Paper NE-89. Upper Darby, PA: US Department of Agriculture, Forest Service, Northeastern Forest Experiment Station.

Fernández-Ugalde, O. and Tóth, G. 2017. Pedotransfer functions for predicting organic carbon in subsurface horizons of European soils. European Journal of Soil Science, **68:** 716-725.

Finke, P.A., Wladis, D., Kros, J., Pebesma, E.J., Reinds, G.J. 1999. Quantification and simulation of errors in categorical data for uncertainty analysis of soil acidification modelling. Geoderma, **93:** 177-194.

Foldal, C., Jandl, R., Bohner, A., Berger, A. 2020. Deriving regional pedotransfer functions to estimate soil bulk density in Austria. Die Bodenkultur: Journal of Land Management, Food and Environment, **71(4):** 241-252.

Friedlingstein, P., O'Sullivan, M., Jones, M.W., Andrew, R.M., Gregor, L., Hauck, J., Le Quéré, C., Luijkx, I.T., Olsen, A., Peters, G.P., Peters, W., Pongratz, J., Schwingschakl, C., Sitch, S., Canadell. J.G., Ciais, P., Jackson, R.B., Alin, S.R., Alkama, R., Arneth, A., Arora, V.K., Bates, N.R., Becker, M., Bellouin, N., Bittig, H.C., Bopp, L., Chevallier, F., Chini, L.P., Cronin, M., Evans, W., Falk, S., Feely, R.A., Gasser, T., Gehlen, M., Gkritzalis, T., Gloege L., Grassi, G., Gruber, N., Gürses, O., Harris, I., Hefner, M., Houghton, R.A., Hurtt, G.C., Iida, Y., Ilyina, T., Jain, A.K., Jersild, A., Kadono, K., Kato, E., Kennedy, D., Goldewijk, K.K., Knauer, J., Korsbakken, J.I., Landschützer, P., Lefèvre, N., Lindsay, K., Liu, J., Liu, Z., Marland, G., Mayot, N., McGrath, M.J., Metzl, N., Monacci, N.M., Munro, D.R., Nakaoka, S.-I., Niwa, Y., O'Brien, K., Ono, T., Palmer, P.I., Pan, N., Pierrot, D., Pocock, K., Poulter, B., Resplandy, L., Robertson, E., Rödenbeck, C., Rodriguez, C., Rosan, T.M., Schwinger, J., Séférian, R., Shutler, J.D., Skjelvan, I., Steinhoff, T., Sun, Q., Sutton, A.J., Sweeney, C., Takao, S., Tanhua, T., Tans, P.P., Tian, X., Tian, H., Tilbrook, B., Tsujino, H., Tubiello, F., van der Werf, G.R., Walker, A.P., Wanninkhof, R., Whitehead, C., Willstrand Wranne, A., Wright, R., Yuan, W., Yue, C., Yue, X., Zaehle, S., Zeng, J., and Zheng, B. 2022. Global Carbon Budget 2022. Earth system science data, **14:** 4811-4900.

Furze, S. and Arp, P.A. 2018. From Soil Surveys to Pedotransfer Function Development and Performance Assessment. *Open Journal of Soil Science* (submitted).

Genuer, R., and Poggi, J.-M. 2020. Introduction to Random Forests with R. In: Random Forests with R. Switzerland: Springer International Publishing AG, pp. 1-8.

Gharahi Ghehi, N., Nemes, A., Verdoot, A., Van Ranst, E., Cornelis, W.M., and Boeckx, P. 2012. Nonparametric techniques for predicting soil bulk density of tropical rainforest topsoils in Rwanda. Soil Sci. Soc. Am. J., **76:** 1172-1183.

Gong, Z., Zhang, X., Chen, J., and Zhang, G. 2003. Origin and development of soil science in ancient China. Geoderma, **115:** 3-13.

Goovaerts, P. 2001. Geostatistical modelling of uncertainty in soil science. Geoderma, **103(1-2):** 3-26.

Gosselink, J.G. and Hatton, R. 1984. Relationship of organic carbon and mineral content to bulk density in Louisiana marsh soils. Soil Science, **137(3):** 177-180.

Gunarathna, M.H.J.P., Sakai, K., Nakandakari, T., Momii, K., Kumari, M.K.N. 2019. Machine learning approaches to develop pedotransfer functions for tropical Sri Lankan soils. Water, **11(1940):** 1-23.

Guo, L., Fan, G., Zhang, Y., Shen, Z. 2019. Estimating the bulk density in 0-20 cm of tilled soils in China's Loess Plateau using support vector machine modeling. Communications in Soil Science and Plant Analysis, **50(14):** 1753-1763.

Hartemink, A.E., and McBratney, A. 2008. A soil science renaissance. Geoderma, **148:** 123-129.

Hastie, T., Tibshirani, R., Friedman, J. 2009. Elements of Statistical Learning, 2nd ed. New York: Springer.

Heuvelink, G.B.M., and Brown, J.D. 2007. Chapter 8. Towards a soil information system for uncertain soil data. In: Lagacherie, P., McBratney, A.B., and Voltz, M. (Eds). Developments in Soil Science, vol. 31. Elsevier, Amsterdam. pp. 91-106

Hikouei, I.S., Christian, J., Kim, S.S., Sutter, L.A., Durham, S.A., Yang, J.J. and Vickery, C.G. 2021. Use of random forest model to identify the relationships among vegetative species, salt marsh soil properties, and interstitial water along the Atlantic coast of Georgia. Infrastructures, **6(70):** 1-13.

Jalabert, S.S.M., Martin, M.P., Renaud, J.-P., Boulonne, L., Jolivet, C., Montanarella, L., and Arrouays, D. 2010. Estimating forest soil bulk density using boosted regression modelling. Soil Use and Management, **26:** 516-528.

Jansen, M.J.W. 1998. Prediction error through modelling concepts and uncertainty from basic data. Nutrient Cycling in Agroecosystems, **50:** 247-253.

Jeffrey, D.W. 1970. A note on the use of ignition loss as a means for the approximate estimation of soil bulk density. Journal of Ecology, **58(1):** 297-299.

Jenny, H. Factors of soil formation: a system of quantitative pedology. McGraw-Hill, New York. 1941.

Johnson, M.L. 2008. Nonlinear least-squares fitting methods. In: Methods in Cell Biology, Vol. 84, pp. 781-805. United States: Elsevier Science and Technology.

Johnson, M.L. and Frasier, S.G. 1985. [16] Nonlinear least squares analysis. Methods in Enzymology, **117:** 301-342.

Kasraei, B., Heung, B., Saurette, D.D., Schmidt, M.G., Bulmer, C.E., Bethel, W. 2021. Quantile regression as a generic approach for estimating uncertainty of digital soil maps produced from machine-learning. Environmental Modelling and Software, **144:** 105139.

Katuwal, S., Knadel, M., Norgaard, T., Moldrup, P., Greve, M.H., and de Jonge, L.W. 2020. Predicting the dry bulk density of soils across Denmark: Comparison of single-parameter, multi-parameter, and vis-NIR based models.

Khodaverdiloo, H., Bahrami, A., Rahmati, M., Vereecken, H., Miryaghoubzadeh, M., Thompson, S. 2022. Recalibration of existing pedotransfer functions to estimate soil bulk density at a regional scale. Eur. J. Soil Sci., **73:** e13244.

Koenker, R. and Bassett, G. 1978. Regression quantiles. Econometrica, **46(1):** 33-50.

Koenker, R. 2017. Quantile regression: 40 years on. Annual Review of Economics, **9:** 155-176.

Krogh, L., Breuning-Madsen, H., and Greve, M.H. 2000. Cation-Exchange capacity pedotransfer functions for Danish soils. Acta Agric. Scand., Sect. B, Soil and Plant Sci., **50:** 1-12.

Krol, B.G.C.M. 2008. Chapter 11. Towards a data quality management framework for digital soil mapping with limited data. In: Lagacherie, P., McBratney, A.B., and Voltz, M. (Eds). Digital Soil Mapping: An Introductory Perspective. Developments in Soil Science, vol. 31. Elsevier, Amsterdam, pp. 137-149.

Lagacherie, P., McBratney, A.B. 2007. Chapter 1. Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. In: Lagacherie, P., McBratney, A.B., and Voltz, M. (Eds). Digital Soil Mapping: An Introductory Perspective. Developments in Soil Science, vol. 31. Elsevier, Amsterdam, pp. 3-24.

Lagacherie, P. 2008. Chapter 1. Digital soil mapping: A state of the art. In: Hartemink, A.E., McBratney, A.B., and Mendonça Santos, M.D.L. (Eds) Digital Soil Mapping with Limited Data. Springer, Netherlands. pp.3-14.

Lal, R. 2004a. Soil carbon sequestration impacts on global climate change and food security. Science, **304(5677):**1623-1627.

Lal, R. 2004b. Soil carbon sequestration to mitigate climate change. Geoderma, **123:** 1-22.

Lal, R. 2018. Soil and Climate. *In* Soil and Climate, Lal, R. and Stewart, B.A. (Eds). CRC Press, Taylor and Francis Group

Levi, M.R. 2017. Modified Centroid for estimating sand, silt, and clay from soil texture class. Soil Sci. Soc. Am. J. **81:** 578-588.

Liao, K., Xu, S., and Zhu, Q. 2015. Development of ensemble pedotransfer functions for cation exchange capacity of soils of Qingdao in China. Soil Use and Management, **31:** 483-490.

Ließ, M., Glaser, B., and Huwe, B. 2012. Uncertainty in the spatial prediction of soil texture: comparison of regression tree and Random Forest models. Geoderma, **170:** 70-79.

Lombardo, L., Saia, S., Schillaci, C., Mai, P.M., and Huser, R. 2018. Modeling soil organic carbon with quantile regression: Dissecting predictors' effects on carbon stocks. Geoderma, **318:** 148-159.

Malone, B.P., McBratney, A.B., Minasny, B. 2011. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. Geoderma, **160:** 614-626.

Martin, M.P., Lo Seen, D., Boulonne, L., Jolivet, C., Nair, K.M., Bourgeon, G., and Arrouays, D. 2009. Optimizing pedotransfer functions for estimating soil bulk density using boosted regression trees. Soil Sci. Soc. Am. J., **73:** 485-493.

McBratney, A.B., Minasny, B., Cattle, S.R., Vervoort, R.W. 2002. From pedotransfer functions to soil inference systems. Geoderma, **109:** 41-73.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B. 2003. On digital soil mapping. Geoderma, **117:** 3-52.

McKeague, J.A. and Stobbe, P.C. History of Soil Survey in Canada 1914 – 1975. Research Branch, Canada Department of Agriculture. Historical Series No. 11, 1978.

Miller, B.A. and Schaetzl, R.J. 2014. The historical role of base maps in soil geography. Geoderma, **230-231:** 329-339.

Minasny, B., and Hartemink, A.E. 2011. Predicting soil properties in the tropics. Earth-Science Reviews, 106: 52-62.

Minasny, B., and McBratney, A.B. 2016. Digital soil mapping: A brief history and some lessons. Geoderma, **264:** 301-311.

Minasny, B., McBratney, A.B., Wadoux, A.M.J.-C., Akoeb, E.N., and Sabrina, T. 2020. Precocious early 19th century soil carbon science. Geoderma Regional, **22:** e00306.

Mwango, S.B., Wickama, J., Msanya, B.M., Kimaro, D.N., Mbogoni, J.D. and Meliyo, J.L. 2019. The use of pedo-transfer functions for estimating soil organic carbon contents in maize cropland ecosystem in the Coastal Plains of Tanzania. Catena **172:** 163-169.

Nanko, K., Ugawa, S., Hashimoto, S., Imaya, A., Kobayashi, M., Sakai, H., Ishizuka, S., Miura, S., Tanaka, N., Takahashi, M., Kaneko, S. 2014. A pedotransfer function for estimating bulk density of forest soil in Japan affected by volcanic ash. Geoderma, **213:** 36-45.

Obidike-Ugwu, E., Ogunwole, J., and Eze, P.N. 2022. Derivation and validation of a pedotransfer function for estimating the bulk density of tropical forest soils. Modeling Earth Systems and Environment, **9:** 801-809.

Omuto, C., Nachtergaele, F., and Vargas Rojas, R. 2013. State of the art report on global and regional soil information: Where are we? Where to go? Food and Agriculture Organization of the United Nations.

Palladino, M., Romano, N., Pasolli, E., Nasta, P. 2022. Developing pedotransfer functions for predicting soil bulk density in Campania. Geoderma, **412:** 115726.

Pribyl, D.W. 2010. A critical review of the conventional SOC to SOM conversion factor. Geoderma, **156:** 75-83.

Ramcharan, A., Hengl, T., Beaudette, D., and Wills, S. 2017. A soil bulk density pedotransfer function based on machine learning: A case study with the NCSS soil characterization database. Soil Sci. Soc. Am. J., **81:** 1279-1287.

Ritz, C. and Streibig, J.C. 2008. Nonlinear Regression with R. 1st ed. New York, NY: Springer.

Saini, G.R. 1966. Organic matter as a measure of bulk density of soil. Nature, **210:** 1295-6.

Sequeira, C.H., Wills, S.A., Seybold, C.A., West, L.T. 2014. Predicting soil bulk density for incomplete databases. Geoderma, **213:** 64-73.

Schillaci, C., Perego, A., Valkama, E., Märker, M., Saia, S., Veronesi, F., Lipani, A., Lombardo, L., Tadiello, T., Gamper, H.A., Tedone, L., Moss, C., Pareja-Serrano, E., Amato, G., Kühl, K., Dămătîrcă, C., Cogato, A., Mzid, N., Eeswaran, R., Rabelo, M., Sperandio, G., Bosino, A., Bufalini, M., Tunçay, T., Ding, J., Fiorentini, M., Tiscornia, G., Conradt, S., Botta, M., Acutis, M. 2021. New pedotransfer approaches to predict soil bukl density using WoSIS soil data and environmental covariates in Mediterranean agro-ecosystems. Science of the Total Environment, **780:** 146609.

Schils, R., 2011. Svante Arrhenius. In *How James Watt Invented the Copier*. New York, NY: Springer New York, pp. 103–109.

Silatsa, F.B.T., Yemefack, M., Tabi, F.O., Heuvelink, G.B.M., Leenaars, J.G.B. 2020. Assessing countrywide soil organic carbon stock using hybrid machine learning modelling and legacy soil data in Cameroon. Geoderma, **367:** 114260.

Sorenson, P.T., Shirtliffe, S.J., Bedard-Haugh, A.K. 2021. Predictive soil mapping using historic bare soil composite imagery and legacy soil survey data. Geoderma, 401: 115316.

Stewart, V.I., Adams, W.A., and Abdulla, H.H. 1979. Quantitative pedological studies on soils derived from Silurian mudstones. II. The relationship between stone content and apparent density of the fine earth. J. Soil Sci., **21:** 248-255.

Sun, W. and Yuan, Y.-X. 2006. Optimization Theory and Methods: Nonlinear Programming. New York, NY: Spring Science + Business Media.

Tranter, G., Minasny, B., McBratney, A.B. 2010. Estimating pedotransfer function prediction limits using fuzzy k-means with extragrades. Soil Sci. Soc. Am. J., **74:** 1967-1975.

Van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., Montzka, C., Nemes, A., Pachepsky, Y.A., Padarian, J., Schaap, M.G., Tóth, B., Verhoef, A., Vanderborght, J., van der Ploeg, M.J., Weihermüller, L., Zacharias, S., Zhang, Y., and Vereecken, H. 2017. Pedotransfer functions in Earth System Science: Challenges and perspectives. Reviews of Geophysics, **55:** 1199-1256.

Van Zijl, G.M., Ellis, F., and Rozanov, A. 2014. Understanding the combined effect of soil properties on gully erosion using quantile regression. South African Journal of Plant and Soil, **31(3):** 163-172.

Vaysse, K. and Lagacherie, P. 2015. Evaluating digital soil mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). Geoderma Regional, **4:** 20-30.

Wadoux, A.M.J.-C., Minasny, B., McBratney, A.B. 2020. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews, **210:** 103359.

Zihao, H., Shaofei, J., and Ku, W. 2022. Application of machine learning methods for estimation soil bulk density. 2022 2nd Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS), Shenyang, China, pp. 194-198.

Zinn, Y.L., Lal, R., and Resck, D.V.S. 2005. Texture and organic carbon relations described by a profile pedotransfer function for Brazilian Cerrado soils. Geoderma **127:** 168-173.

Chapter 2. A Framework for Recalibrating Pedotransfer Functions using Nonlinear Least Squares and Estimating Uncertainty using Quantile Regression

A version of this chapter has been published in the journal Geoderma:

Arbor, A., Schmidt, M., Saurette, D., Zhang, J., Bulmer, C., Filatow, D., Kasraei, B., Smukler, S., Heung, B. 2023. A framework for recalibrating pedotransfer functions using nonlinear least squares and estimating uncertainty using quantile regression. Geoderma, **439:** 116674.

2.1. Abstract

Pedotransfer functions (PTFs) have been developed for many regions to estimate values missing from soil profile databases. However, globally there are many areas without existing PTFs, and it is not advisable to use PTFs outside their domain of development due to poor performance. Further, developed PTFs often lack accompanying uncertainty estimations. To address these issues, a framework is proposed where existing equation-based PTFs are recalibrated using a nonlinear least squares (NLS) approach and validated on two regions of Canada; this process is coupled with the use of quantile regression (QR) to generate uncertainty estimates. Many PTFs have been developed to predict soil bulk density, so this variable is used as a case study to evaluate the outcome of recalibration. New coefficients are generated for existing soil bulk density PTFs, and the performance of these PTFs is validated using three case study datasets, one from the Ottawa region of Ontario and two from the province of British Columbia, Canada. The improvement of the performance of the recalibrated PTFs is evaluated using root mean square error (RMSE) and the concordance correlation coefficient (CCC). Uncertainty estimates produced using QR are communicated through the mean prediction interval (MPI) and prediction interval coverage probability (PICP) graphs. This framework produces dataset-specific PTFs with improved accuracy and minimized uncertainty, and the method can be applied to other regional datasets to improve the estimations of existing PTF model forms. The methods are most successful with large datasets and PTFs with fewer variables and minimal transformations; further, PTFs with organic carbon (OC) as one of or the sole input variable resulted in the highest accuracy.

2.2. Introduction

Soil data are an invaluable resource, and the demand for soil data is increasing. Soil data are used for monitoring soil health, functions, and nutrient fluxes and storage. The need for accessible, standardised global soil databases has long been identified (Batjes et al., 1994), and soil data repositories around the world have been growing as records and maps have been digitised (Arrouays et al., 2017). Soil data are a prime source of input for data repositories from local- to global-scales, however, they are often incomplete. Pedotransfer functions (PTFs) have been used to estimate values missing from soil profile databases (Baritz et al., 2010; Benites et al., 2007; Tranter et al., 2007). PTFs are quantitative functions that explain relationships between soil variables, which "translate the data we have to the data we need" (Bouma,1989) — they are numerous and have been developed for many soil attributes, especially soil hydraulic properties and bulk density. While PTFs are useful tools, their application comes with several issues. Firstly, there are many PTFs available to choose from—this is especially true for bulk density, with papers cataloguing up to 63 bulk density PTFs tested on a dataset and compared (Nasta et al., 2020). Secondly, a selected model generated from another geographical region may not be transferrable to a new region (McBratney et al., 2002). Lastly, PTFs often do not have accompanying uncertainty estimates—a concern which has previously been identified (McBratney et al., 2002).

With existing soil datasets, filling in the missing values required for a particular study requires selecting an existing PTF or developing a new one; a third option is to recalibrate an existing PTF. PTFs are continuously being developed for regions for which few or none exist, such as for tropical soils (Obidike-Ugwu et al., 2023). Selecting an existing PTF may be difficult, as it has been argued that PTFs should not be transferred from outside their region of development (Casanova et al., 2016; De Vos et al., 2005; McBratney et al., 2002; Van Looy et al., 2017). Recalibrating existing equation-based PTFs is an alternative to developing a new PTF and has the advantage that it does not require that the dataset used to calibrate the original function be similar to the dataset on which it is being recalibrated. Recalibration has been carried out in previous studies; for example, De Vos et al. (2005) recalibrated two bulk density PTFs; Reidy et al. (2016) used multiple regression analysis to recalibrate functions; Nanko et al. (2014)

and Chen et al. (2018) used Levenberg-Marquardt (Marquardt, 1963) nonlinear least squares (NLS) to revise existing PTFs, and to recalibrate six model forms, respectively.

Estimates of uncertainty are rarely provided with developed PTFs (McBratney et al., 2002), and this has been noted as a limitation of many studies which use or develop PTFs (Nemes, 2015). Uncertainty estimates should be provided with a PTF, and the one with the smallest error variance should be chosen (McBratney et al., 2002; Minasny and Hartemink, 2011). The provision of uncertainty estimates ensures that the data they generate are used with appropriate care, particularly in the context of management of policy decisions. One method of estimating uncertainty uses quantile regression (QR; Koenker and Bassett, 1978). The QR approach has been used for a variety of applications (Muthusamy et al., 2016; Rahmati et al., 2019; López López et al., 2014); in soil science, Kasraei et al. (2021) developed a generic framework for estimating the uncertainty of digital soil maps produced from machine-learning techniques, using QR.

While there has been much work invested in developing PTFs, there is a need for establishing clear protocols for adapting PTF equations to regional datasets and generating their uncertainty. This study focusses on PTFs generated for bulk density as a case study due to their wide availability, as well as their importance in estimating soil organic carbon (SOC) stocks and hydrological soil properties. Bulk density may be measured through multiple methods, which include excavation, the core and the clod methods (Al-Shammary et al., 2018); however, bulk density values are often missing or have limited availability in soil databases. To estimate missing values, approaches include regression and machine learning (Minasny et al., 1999). Hence, the goals of this study are (1) to provide a framework to recalibrate and validate existing PTF model forms for local datasets using the NLS approach to generate new model coefficients; (2) to compare the predictive performance of the PTFs generated using NLS to both the original PTFs found within the literature, and to other recalibrated PTFs in order to identify the best performing model form; and (3) to generate and evaluate the performance of uncertainty estimation using QR. Here, we apply a framework to expand the coverage of PTFs for regions or conditions for which no PTFs exist by using the Province of British Columbia (BC) and the Ottawa region of Ontario, Canada, as case study areas.

2.3. Methodology

Two case study regions were considered in this study: the Province of BC, and the Ottawa region of southern Ontario, Canada. Three datasets were developed to test PTFs: Ontario (All Variables), BC (All Variables), and BC (Carbon and Bulk Density). A literature search for existing, equation-based PTFs targeting bulk density was conducted. Two datasets, Ontario (All Variables) and BC (All Variables), were tested on all PTFs obtained from the literature search. The BC (Carbon and Bulk Density) dataset was tested with functions that only contained SOC or organic matter (OM) as an input variable. Model forms common to multiple PTFs were identified, and PTFs were divided into 8 groups; these groups were based on both the form of the equation and the input variables utilized. Following this, recalibration of the functions was performed using the NLS approach, and new coefficients for the PTFs were generated. The results of the NLS approach were compared to the performances of the PTFs in their original literature coefficients, and to other recalibrated PTFs of varying model forms and input variables. Simultaneously, QR was performed, and the accuracy and uncertainty metrics generated for the PTFs.

2.3.1. Datasets

The datasets from BC and Ontario contained data from mineral soil horizons with SOC values < 17% (by weight). The data was limited to mineral soil with the anticipated result that the PTFs could be better fit to a dataset that contained a smaller range of values: in this case, a restricted range of SOC values. BD shows an inverse relationship with SOC, and organic horizons with high SOC values have very low BD values. For the BC data, organic soil had a mean BD of 0.38 g/cm^3 and SOC of 39.7%, while the mineral soil had a mean BD of 1.33 g/cm^3 and SOC of 2.19%. For the Ontario data, organic soils had a mean BD of 0.24 g/cm^3 and SOC of 34.1%, while mineral soils had a mean BD of 1.17 g/cm^3 and SOC of 2.29%. Some previous studies which tested or developed PTFs have focused on doing so for specific soil conditions, such as Nanko et al. (2014), which examined volcanic influenced forest soils in Japan; or have partitioned datasets based on attributes such as depth (Kätterer et al., 2006), parent materials (Heinonen, 1977), cultivation status (Hollis et al., 2012), mineral or organic (Hossain et al., 2015), taxonomy (Saini, 1966; Alexander, 1989), the presence of a particle size in certain quantities, such

as clay (Beutler et al., 2017) or sand (Bernoux et al., 1998); or other environmental variables. Partitioning a dataset allows a PTF to be developed for a certain condition, which has the potential result of increasing the accuracy of the PTF.

Soil Datasets

Attributes, such as bulk density, SOC, depth (calculated as midpoint of the horizon), pH (H_2O), coarse fragment content, silt, sand, and clay were included in the Ontario (All Variables) and the BC (All Variables) datasets. The BC (Carbon and Bulk Density) dataset only contained SOC and bulk density as attributes, to maximise the number of points that could be included in the dataset. This dataset included the horizons contained in the BC (All Variables) dataset, plus an additional 803 horizons; these additional horizons had observations of BD and SOC, but did not have the full set of attributes required for inclusion in the BC (All Variables) dataset. Summary statistics for all datasets are available in Table 2.1. No bulk density points that were < 0.3 g/cm^3 were included; it was assumed that it would be unlikely for any mineral soil to have a bulk density value that low. In other studies, (Périé and Ouimet, 2008; Tremblay et al., 2002; Federer et al., 1993), for context, the bulk density of pure OM was estimated to be 0.11 – 0.12 g/cm^3. Honeysett and Ratkowsky (1989) estimated pure OM to have a bulk density of 0.163 g/cm^3, and Adams (1973) found a range of 0.207 to 0.291 g/cm^3, depending on moisture state. As a mineral soil contains < 17% organic carbon (OC), rather than 58% OC in pure OM (using the traditional van Bemmelen conversion factor), the minimum bulk density value for a mineral soil should therefore be higher. Abdelbaki (2018) removed any sample whose bulk density was < 0.30 g/cm^3; and other studies, which based their PTFs on mineral soils (rather than on depth interval or other specification), had minimum bulk density values in a similar range: 0.35 g/cm^3 (Barros and Fearnside, 2015), 0.35 g/cm^3 (Grigal et al., 1989), and 0.31 g/cm^3 (Hollis et al., 2012). Further, the Canadian System of Soil Classification (Soil Classification Working Group, 1998), provides ranges for BD values for fibric, mesic, and humic materials found in the O horizons (i.e., Of, Om, and Oh). These have been found to be very low, with the BD of fibric material < 0.075 g/cm^3. The BC datasets did not contain any bulk density values < 0.3 g/cm^3, so no points were removed from the datasets on that basis. The Ontario (All Variables) dataset did contain 16 horizons with bulk density values < 0.3 g/cm^3, which were removed.

British Columbia Case Study

The complete BC dataset contains points that occur at various locations in the province of BC. The province extends from 48°17' 52.9" N in the south to 60°00'00" N, the Yukon border; and from 114°04'00" W at the Alberta border to 139°03'00" W. Multiple north-south mountain ranges cross the province, and elevation ranges from 0 m above sea level to the 4,663 m summit of Mount Fairweather on the Alaskan border. The province has an area of 944,735 km^2, and encompasses a wide range of topography, geology, climate, and vegetation; as a result of this diversity, the soils of BC are also diverse, and all soil orders defined by the Canadian System of Soil Classification (CSSC) are found in BC.

All data were acquired from the BC Soil Information System (BCSIS) datasets. For BC (All Variables), there were 396 horizons, from 129 sites, for a density of 7,324 km^2 per site; BC (Carbon and Bulk Density) had 1,199 horizons, from 441 sites, for a density of 2,142 km^2 per site. Conversion factors were applied to several variables, to maximise the number of available points. pH in $CaCl_2$ values were correlated to pH in H_2O using a linear regression with an R^2 of 0.73. The pH ($CaCl_2$) values were then converted using the equation:

$$pH_{(H2O)} = 0.9757 * pH_{(CaCl2)} + 0.7143 \quad (1)$$

The use of the conversion factor does introduce error into the measurement; this error could then be propagated through the bulk density estimations made using converted values. However, only a small percentage of pH values were calculated using this conversion factor, and the presence of these values allowed the inclusion of more horizons in the dataset. Bulk density values were acquired from BCSIS with no conversion between analytical methods, which included the following four methods: volumeter, saran (also known as the clod method), excavation, and core (Blake, 1965). Three lab methods were used to analyse carbon: LECO (Wang and Anderson, 1998), Walkley-Black (Walkley and Black, 1934), and loss on ignition (LOI; Ball, 1964; Skjemstad and Baldock, 2007). Carbon was analysed using LOI for only 12 horizons, and these were not included because the conversion factor was developed on a small sample size, which would introduce additional error. Only LECO and Walkley-Black values were used, with the determined OM values already converted and presented as OC in the dataset, so no additional conversion was necessary. Sand, silt, and clay were

measured using either the hydrometer method or the pipette method (Kroetsch and Wang, 2007) and expressed as percentages, so no conversion was necessary. Coarse fragment content was defined as the gravel content, material that that is between 2.0 and 80 mm in diameter (Kroetsch and Wang, 2007). It was expressed as the unit mass of coarse fragments per unit mass of bulk soil. Depth (cm) was calculated as the midpoint between the top and bottom depth values for each horizon or sample depth interval. Sample site locations for the BC (All Variables) dataset and the BC (Carbon and Bulk Density) dataset are shown in Figure 1.

Ottawa Case Study

The Ontario dataset covers a 2,824 km^2 region near Ottawa, Ontario, extending between 44°57’43” N to 45°32’02” N, and from 75°14’45” W to 76°21’20” W, in the Mixedwood Plains Ecozone of southern Ontario (Figure 2.1). Agricultural uses dominate, with 57% of the land cover classified as cropland, pasture, and abandoned fields. Forests are predominantly deciduous, although coniferous and mixed forest are represented in the 30.1% of forested area (Crins et al., 2009). Elevation differences in the study area are modest, ranging from 55 to 171 m above sea level. Most soils in the area are classified as Orthic Humic Gleysols or Orthic Melanic Brunisols (Soil Classification Working Group, 1998). The density of points was much higher than for the BC dataset. The Ontario dataset contains 3,242 horizons from 2,110 sites, for a density of 1.3 km^2 per site.

There were no analytical conversions necessary for the Ontario dataset. Both pH (H_2O) and pH ($CaCl_2$) were available for all horizons; however, pH (H_2O) was used as this was also the method used in the BC datasets and because the PTFs being tested, which included pH as an input variable, used this method as well. Both coarse fragment content (field estimated) and gravel content (determined in laboratory as part of particle size analysis) values were available. If a horizon had both attributes available, the highest of the two was chosen. All SOC values were determined using the LECO method and hence, an analytical conversion factor was not needed. Samples with extremely low values of SOC were removed (<0.01%), and typically associated with C horizons, or parent material. Sample depth (cm) was represented as the midpoint between the top and bottom of the sampled horizons.

2.3.2. Selecting PTFs and Classification into Model Forms

Previous studies have evaluated and compared existing literature PTFs; studies relevant to this research are presented in Table 2.2. For this study, PTFs for bulk density whose input attributes were contained in the datasets used in the current study were selected from the literature. The input attributes included SOC, coarse fragments, pH, sand, silt, clay, and depth. Other studies have incorporated other variables such as electrical conductivity and $CaCO_3$ concentration (Alaboz, 2020), or environmental variables (Akpa, 2018); for example, Schillaci et al. (2021) incorporated bioclimatic and topographic covariates and found that their inclusion improved model accuracy. However, these variables were not available for the datasets used as case studies. Further, environmental variables were not used because this study was focused on recalibrating PTFs using only the intrinsic properties of soil and whereby the predicted bulk density data could subsequently be used to generate digital soil maps using soil-environmental variables in future studies. In total, 73 PTFs from the literature were selected and are presented in Table 2.3; each PTF was assigned a unique identifier number for the purposes of this study. The original form of each PTF was included, as well as the soil type or region for which the PTF was developed; other information included was the units used in the original function, the original results of evaluation metrics used in the study, and the sample size used to develop the PTF.

Model Types

Recurring model forms were identified within the 73 PTFs selected for this study, and the PTFs were divided into groups based on their model form (Table 2.4). Some model forms have been used in multiple PTFs with varying coefficients, while others were unique. Previous studies, such as De Vos et al. (2005), Nanko et al. (2014) and Chen et al. (2017) have identified and categorized model forms. For example, De Vos et al. (2005) classified the PTFs tested in their study into five model groups: those with log-transformed variables; square root-transformed variables; the reciprocal of bulk density; log-transformed variables and second order polynomials; and multiple predictor variables. Similarly, Nanko et al. (2014) classified the PTFs into six model types: the Stewart/Adams physical model (Adams, 1973) that used the bulk densities of the pure organic and mineral fractions; those with radical root variables; logarithmic-transformed variables; exponential variables; decimal variables; and polynomial variables. Chen et al.

(2017) identified four model types: the Adams equation (Adams, 1973); “logarithmic or exponential functions of SOM”; “radical functions of SOM”; and “functions of multiple variables associated with soil physiochemical properties”. In this study, we categorized PTFs based on the form of the model and the input variables used. Group A contains linear functions with OC or OM as an input. Group B contains radical root-transformed values of OC. Group C contains reciprocal equations. Group D is based on Curtis and Post’s (1964) equation with multiple natural log terms. Group E uses the natural exponent with an intercept term; the natural exponent multiplied by a term; or a natural exponent with a root as well. Group F contains functions with only OC or OM as input variables, but which do not fall into any of the above categories. Group G functions have OC or OM and any of the other attributes (sand, silt, clay, depth, pH, coarse fragment) as input variables; and Group H functions contain any input variables except OC or OM. Lastly, functions in Group X were not successfully recalibrated with NLS.

One model type that has been used multiple times but was not included as it was based on inputs of the bulk densities of pure OM and pure mineral soil (Adams, 1973; Federer, 1993; Nanko, 2014; Prévost, 2004; and Tranter, 2007). It has been referred to as a “physical equation” (Nanko, 2014), and the “organic density approach” (Prévost, 2004). The equation takes the form:

$$BD = \frac{100}{\frac{OM}{BD_O} + \frac{100 - OM}{BD_M}} \quad (2)$$

where OM is the percent organic matter; BDO is the bulk density of pure organic matter; and BDM is the bulk density of pure mineral matter. The bulk densities it requires are measurements which must be experimentally determined for the specific dataset for which the PTF is to be used. As this study utilized legacy data, it was not possible to determine pure bulk densities for OM and mineral soil. Furthermore, as these PTFs do not contain coefficients, which can be recalibrated through NLS, they were not included in the study.

2.3.3. Nonlinear Least Squares

NLS is an optimisation method to estimate new model coefficients. The Gauss-Newton method of solving NLS problems is an iterative method that locates the

minimizer of the function which is a weighted sum of squared terms (Teunissen, 1990). The iterative technique uses an initial estimate of a solution, and then generates a further sequence of estimates based on predetermined rules, with the goal of converging on the solution (Teunissen, 1990). A model form is specified, along with starting values for each coefficient, the starting values used in this study being the original literature coefficients for the PTF. The stats package within the R statistical software (R Core Team, 2022) includes an NLS function, which uses the Gauss-Newton algorithm by default. This function iteratively improves estimations for the coefficients (Bates and Watts, 1988). Coefficient values generated in each of the 200 loops were then averaged, and these averaged values were used as the final coefficients for the model. For the PTFs that used OM as an input variable, the van Bemmelen conversion factor of 1.724 was used to convert the OC values in the dataset to OM so that the output coefficients and evaluation metrics could be compared to the original literature coefficients. For this reason, PTFs equation forms that used OM and OC as input variables were listed with coefficients for both. Although the original coefficients used in the literature PTFs varied in the number of decimal places used, to be consistent, three decimal places were chosen for the NLS generated coefficients.

Assessment of Recalibrated PTFs

All PTFs were cross-validated using repeated nested k-fold cross-validation, whereby the validation statistics were based on the outer loop of the nested cross-validation (see Figure 2.2 for a schematic of the process). While using a separate dataset to validate the results is ideal, dividing the dataset into training/calibration and validation is the more typical approach as acquiring soil data is expensive (McBratney et al., 2011). To minimise autocorrelation between soil measurements made from the same soil profile, model validation was carried out using leave-profile-out cross-validation to ensure that the accuracy metrics were not compromised.

To assess the accuracy of the PTF estimations, two metrics were included: root mean square error (RMSE), and the concordance correlation coefficient (CCC). RMSE is a measurement of fit of a predictive model to a dataset. It calculates the average distance between the predicted values and the observed values, with a lower RMSE indicating a better fit to the data. It is calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{n}} \quad (3)$$

where x_i is the observed value, and $\hat{x}_i$ is the predicted value, of the i^{th} observation, with n being the number of observations. The CCC is a measurement of the agreement between two variables, the predicted values and the observed values. A higher CCC value indicates greater agreement between the variables. It is calculated as follows:

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (4)$$

where μ_x and μ_y are the means of the observed and predicted values; σ_x and σ_y are the corresponding variances.

2.3.4. Quantile Regression

To generate uncertainty estimates for PTFs we used QR, which is an alternative to the least squares estimator for linear models (Koenker and Bassett, 1978). An advantage of QR is that it does not assume a parametric distribution of the data and hence, it can be used with data that has a heterogeneous distribution (Cade and Noon, 2003); furthermore, it is also robust to outliers (Lopez Lopez et al., 2014). Another advantage of the QR method is that it considers every source of uncertainty, rather than only one source as most other methods do (Rahmati et al., 2019).

When using QR to estimate uncertainty, a linear relationship between the observed and predicted value of a target variable is assumed. Linear regression models are developed for each specified quantile, with the response variable being the selected quantile of the variable's conditional distribution. A framework for applying QR follows that of López López et al. (2014), Dogulu et al. (2015), Rahmati et al. (2019) and Kasraei et al. (2021). A linear relationship is assumed between observed values (y) and predicted values ($\hat{y}$) for each quantile (τ):

$$y = a_\tau \hat{y} + b_\tau \quad (5)$$

where a_τ is the slope and b_τ is the intercept of the linear regression. To determine a_τ and b_τ, the sum of residuals is minimized in the following loss function:

$$min \sum_{j=1}^{J} \rho_\tau \left(y_j - \left(a_\tau \hat{y}_j + \ b_\tau\right)\right) \quad (6)$$

where y_j and $\hat{y}_j$ are the *j*th paired samples, *J* is the total number of samples, and ρ_τ is the QR function for the quantile τ:

$$\rho_\tau \left(\epsilon_j\right) = \begin{cases} (\tau - 1)\epsilon_j, & \epsilon_j < 0 \\ \tau\epsilon_j, & \epsilon_j > 0 \end{cases} \quad (7)$$

where ϵ_j are the model residuals, calculated as the difference between the observed and predicted values from Eq (1), for the τ^{th} quantile. For the desired quantile τ, the QR function is applied to the residual, ϵ_j in Eq (3).

Readers should note that a full description of QR and its implementation within R is provided in Kasraei et al. (2021), and includes a methodological framework for integrating QR into predictive modelling (see Section 2: Theoretical Background). In Kasraei et al. (2021), the framework was used for generating uncertainty estimates for digital soil maps using machine learning; however, by replacing the machine learner with either a PTF from the literature or a refitted PTF using NLS, uncertainty estimates may be generated for those PTFs. Here, it is important to note that the framework consists of an 'inner' and 'outer' loop to ensure that a matrix of the observed and predicted values can be generated for QR within the 'inner' loop; and that the quality of the uncertainty estimates can be assessed using a fully independent dataset in the 'outer' loop – effectively forming a nested cross-validation procedure. When generating the uncertainty estimates for the literature-based PTFs, the 'inner' loop does not require a model to be fitted to the calibration dataset as model coefficients have already been determined *a priori* from the literature. Within the repeated nested cross-validation procedure, the 'inner' loop uses 10-fold cross-validation and the 'outer' loop is repeated 20 times. The repeats were used to ensure stability in the accuracy metrics. Coefficient values are generated for each iteration, and then are averaged to calculate the new coefficients for the respective PTF equations. Accuracy metrics, such as RMSE and CCC, are also calculated from the outer loop.

2.3.5. Uncertainty Assessment

Quantile regression can be used to calculate mean prediction interval (MPI) and PICP, which can be displayed graphically (see Appendix B). A prediction interval is the range of values which future predicted values are expected to fall within, with a prescribed probability, and it is bounded by upper and lower limits (Shrestha and Solomatine, 2006). Prediction intervals are defined for selected confidence levels, with PICP then calculated for each confidence level as follows (Kasraei et al., 2021):

$$\text{PICP} = \frac{1}{n}\sum_{t=1}^{n} C,\ C = \begin{cases} 1, PL_t^{lower} \leq y_t \leq PL_t^{upper} \\ 0, otherwise \end{cases} \quad (8)$$

where y_t is the observed value, PL_t^{upper} is the upper limit and PL_t^{lower} is the lower limit. To evaluate the PICP results, it is plotted against the corresponding confidence level. A 1:1 relationship between the values is the desired outcome meaning that the predicted values are within the prediction interval (Malone et al., 2011). If the PICP is less than or greater than the confidence level, the uncertainty has been under- or over-estimated, respectively.

To calculate the MPI, the widths of the prediction intervals are averaged (Rahmati et al., 2019). Narrower MPI widths indicate lower uncertainty, and wider MPI widths indicate higher uncertainty (Kasraei, 2021; Rahmati, 2019).

$$\text{MPI} = \frac{1}{n}\sum_{\tau=1}^{n}(PL_\tau^{upper\ limit} - PL_\tau^{lower\ limit}) \quad (9)$$

where $PL_\tau^{upper\ limit}$ is the upper limit and $PL_\tau^{lower\ limit}$ is the lower limit of the PI.

2.4. Results & Discussion

2.4.1. Accuracy Assessment

In nearly all cases, NLS was able to generate coefficients that improved the performance of the PTF. For example, with the Ontario (All Variables) dataset, the average RMSE after recalibration was reduced to 0.19 g/cm^3, from an initial range of 0.19 to 12.48 g/cm^3; the average CCC increased to 0.63, from an initial range of -0.09 to 0.68; and the average R^2 value increased to 0.43 from a range of 0.001 to 0.51.

Recalibration with NLS was most successful with the Ontario dataset. This was likely due to it being the largest dataset, with 3224 observations, and a far higher sampling density. Table 2.5 shows the new coefficients generated for each model form, for each dataset. Table 2.6 contains the performance results, measured in RMSE and CCC, of the original PTFs compared to the PTFs recalibrated with NLS coefficients for Ontario (All Variables); Table 2.7 contains the same comparison but for BC (All Variables); and Table 2.8 contains the results for BC (Carbon and Bulk Density). Figure 2.3 shows an example PTF from Group D, with the observed vs predicted values both before and after recalibration plotted. Figures 2.12 to 2.15 are plots of example PTFs for each dataset, showing Groups A, B, C and E.

It is important to note that of the 73 publications that were reviewed, only 13 publications explicitly stated that an external test dataset was used to generate accuracy metrics while the remaining 60 publications either reported only model fit statistics or were not clear in what their accuracy metrics were representing (Table 2.3). Given that these PTFs are intended to predict bulk density, there is the need to ensure that external test data is used to generate meaningful accuracy metrics and that users of bulk density PTFs should be vigilant when interpreting those metrics when selecting an appropriate PTF.

Preferred model type by dataset

For the Ontario (All Variables) dataset, the model types with the highest accuracy were Groups B, C and E. These were the radical root functions with OC/OM, the reciprocal functions with OC/OM, and functions with natural exponent terms using only OC/OM as an input variable. Recalibration resulted in CCC values of 0.67 and RMSE values of 0.19 g/cm^3 for Group B functions. For Group C PTFs, initial starting CCC values ranged from 0.33 to 0.58, and initial RMSE values from 0.34 to 0.25 g/cm^3. After recalibration, all functions had a resulting CCC of 0.67 and RMSE of 0.19 g/cm^3. Group E PTFs had a greater diversity of model forms, and variously included an intercept term, a root transformation of the OC/OM input, or additional coefficients. Testing of the PTFs with their literature coefficients showed a range of results; CCC values were as low as 0.02, and RMSE values as high as 2.45 g/cm^3. With recalibration, all functions improved: the resulting CCC was either 0.66 or 0.67, and RMSE values were all 0.19 g/cm^3.

PTFs from Groups C, D and E produced estimates with the greatest accuracy after recalibration on the BC (All Variables) dataset. Results for Group C functions had average CCC values of 0.57, and RMSE of 0.29 g/cm^3 after applying NLS. Group D models used natural log transformations of OC/OM, and performed poorly when tested with their original coefficients. This model form is a function which showed large fluctuations in its form when the OC/OM values approached zero; and therefore, the portion of the function used to fit a curve to the data behaved drastically differently when the input data range was altered. Group D functions improved substantially with recalibration, such as PTF-22 (Federer 1983), whose CCC improved from 0.02 to 0.57 and RMSE improved from 1.28 to 0.28 g/cm^3. For Group E functions, some PTFs showed large improvements, such as PTF-37 where an initial CCC of 0.12 increased to 0.57, and the RMSE decreased from 1.10 to 0.29 g/cm^3; or PTF-63, where an initial CCC of 0.03 increased to 0.52, and RMSE decreased from 0.61 to 0.25 g/cm^3.

For the BC (Carbon and Bulk Density) dataset, it was Groups D and E whose estimates had the highest accuracy. Large improvements in accuracy were shown with recalibration with Group D functions; for example, the RMSE of PTF-22 (Federer 1983) decreased from 1.29 to 0.30 g/cm^3, and its CCC increased from 0.03 to 0.54. The recalibrated PTFs from Group E had CCC values ranging from 0.44 – 0.54 and RMSE values of 0.30 – 0.33 g/cm^3.

PTFs from Group A (linear functions) also showed good recalibration results across datasets, although their accuracy was less than the previously mentioned model types. For example, all were recalibrated to an average CCC of 0.62 and RMSE of 0.20 g/cm^3 on the Ontario (All Variables) dataset. Certain PTFs that were not part of the previously discussed model groups but nonetheless showed high accuracy results across datasets include PTF-68 (Sevastas et al., 2018), PTF-14 (Beutler et al., 2017), PTF-31 and PTF-33 (Hollis et al., 2012), and PTF-50 (Kaur et al., 2002). For an overview of how the RMSE and CCC of each PTF differed before and after recalibration, see Figures 2.6 to 2.11. For each dataset, a graph of RMSE values and CCC values is available with every PTF included.

Comparison to results from other studies

Other studies have also tested a variety of PTFs and reported their results; Table 2.9 shows a selection of studies which have tested some of the same PTFs included in

this study and reported their RMSE values. Many original studies only reported the R^2 value, so the RMSE results from other datasets are especially useful for comparison. PTFs which have been tested by many other studies but which only report R^2 values in the original study include PTF-5, PTF-6, PTF-22, PTF-39, PTF-19, PTF-55, and PTF-72.

PTF-50 (Kaur et al., 2002) has been tested multiple times by other studies; RMSE values obtained range from 0.26 to 0.56 g/cm^3. This compares to 0.15 g/cm^3 obtained in the original study; when recalibrated, produced 0.19 g/cm^3 on Ontario (All Variables) and 0.32 on BC (All Variables). PTF-31 and PTF-33 (both Hollis et al., 2012) also had low RMSE values reported by the original study, at 0.13 and 0.15 g/cm^3 respectively. After recalibration, these PTFs both showed an accuracy of 0.19 g/cm^3 on the Ontario (All Variables) dataset and 0.28 g/cm^3 on the BC (All Variables) dataset. PTF-14 (Beutler et al., 2017) had an RMSE of 0.22 g/cm^3 when developed for soils in Brazil; recalibration on two Canadian datasets generated an RMSE range of 0.19 g/cm^3 on the Ontario (All Variables) to 0.29 g/cm^3 on the BC (All Variables) dataset. PTF-68 (Sevastas et al., 2018) had an original study RMSE of 0.12 g/cm^3; recalibrated on Ontario (All Variables) it was 0.19 g/cm^3, on BC (All Variables) 0.29 g/cm^3, and on BC (Carbon and Bulk Density) 0.32 g/cm^3. Of any PTF tested in this study, PTF-60 (Qiao et al., 2019) had the lowest RMSE in its original study, of 0.079 g/cm^3. When recalibrated, the accuracy on Ontario (All Variables) was 0.24 g/cm^3, and 0.33 g/cm^3 on the BC (All Variables) dataset.

2.4.2. Uncertainty Assessment

Uncertainty estimates for PTFs were generated using a QR approach. Graphs of the PICP with respect to the CL were also generated, and a representative PTF from Group D is shown in Figure 2.4; representatives from Groups A, B, C, and E are plotted for each dataset in Figures 2.17 to 2.20. The PICP plots suggest that the greatest factor influencing PICP is the size of the dataset. Within each dataset, PICP plots were nearly indistinguishable for the various PTFs; however, for a given function, its PICP plot varied depending on the dataset used. The example given in Figure 2.4 (PTF-19, Curtis & Post 1964), was typical of the variation between datasets for the same PTF. When the function was recalibrated on the Ontario dataset (with 3,243 horizons), the resulting PICP plot showed estimates tightly grouped on the 1:1 line, indicating low uncertainty. When the BC (All Variables) dataset was used, estimates had a greater spread both

above and below the 1:1 line, indicating greater uncertainty; this also corresponds to the BC (All Variables) dataset being the smallest, with only 396 horizons. The BC (C and BD) dataset is larger, with 1,199 horizons, and the PICP plots reflect lower uncertainty for PTFs tested on that dataset. These results were consistent for all PTFs tested, with the lowest uncertainty estimates for those tested on the Ontario (All Variables) dataset; increased uncertainty for functions tested on the BC (Carbon and Bulk Density) dataset, and the greatest uncertainty for functions tested on the smallest dataset, BC (All Variables).

MPI values were generated, and a comparison of model forms on different datasets was produced, with the MPI for each confidence level displayed in the graphs. Figure 2.5 compares recalibrated functions from Groups A to D for all three datasets; Figure 2.21 shows recalibrated PTFs from Group E; Model forms in Group F are displayed in Figure 2.22. Group G results are shown in Figure 2.23. The results for the final group of recalibrated model forms, Group H, is shown in Figure 2.24.

While the datasets show similar patterns of which model forms had greater or lesser uncertainty associated with their predictions, when the absolute values of MPI are compared across datasets, the Ontario (All Variables) dataset showed the lowest MPI values, indicating the lowest uncertainty. For example, when Groups A to D are considered, the highest MPI value for the 99% CL is 1.13 on the Ontario (All Variables) dataset for the Group A model, but 1.83 on the BC (C and BD) dataset. These findings correspond to the PICP results, which showed the PICP values for Ontario (All Variables) fell in a narrower range than for BC (All Variables), which had the largest range of values. This indicates that lower uncertainty is associated with a larger dataset such as the Ontario (All Variables); the greatest uncertainty occurs when the PTFs are recalibrated on the smallest dataset, the BC (All Variables).

2.4.3. Challenges with NLS

Issues with using NLS to generate new coefficients included sensitivity to starting values; functions with many terms or transformations to terms; certain functions whose behaviour changes drastically when input values are different than the literature values (or approach zero); and dataset size.

For this study, the literature coefficients were used as the starting values when using NLS. As NLS is sensitive to starting values, when the literature coefficient was too far from the solution value for the dataset, NLS would terminate before convergence. For example, PTF-39 produced an error, and the algorithm terminates before coefficients can be generated if the a-coefficient is negative, as it is in the literature; by changing the sign of the a-coefficient to positive, the function can then be recalibrated. The a-coefficient of PTF-39 is quite different from the a-coefficient of other PTFs in the same group (Group D), which led to termination of the algorithm when this PTF was tested on all datasets. There are multiple approaches to generate starting values if suitable initial coefficients cannot be determined or if literature values cause recalibration to fail; these include knowledgeable guesses, trying out multiple values (brute force), plotting the data and fitting a rough curve, or utilizing one of the available self-start functions (Ritz and Streibig, 2008).

Functions with many terms or transformations to terms, such as log or exponent transformations, more frequently produce errors when NLS is attempted; this was the case with PTF-57 (Pereira et al., 2016). The failure to recalibrate for PTF-57 was likely due to the large number of coefficients in the equation, as well as the nested sin transformations of the variables. A solution which corrects most failures to recalibrate is to transform the variables before using NLS. This allows inspection of the transformed variables before use; for example, log transformed variables which have an initial value of zero will produce a null value that causes recalibration to fail. These values can then be identified and changed to zero before using the NLS algorithm. Further, some PTFs needed to be simplified, such as with the Ruehlmann & Körschens (PTFs-61, 62, 63), as recalibration did not work when all the terms were assigned as coefficients. Nguyen (2020) also cites a model that is too complex as a reason that the algorithm could fail to reach convergence.

Functions which have drastic behaviour differences such as those in Group D have varying recalibration results; Group D functions had multiple natural log terms, and only OC or OM as input variables. When this function is graphed using original literature coefficients, it is notable that the behaviour of the function shows extreme changes in the bulk density values it produces when OC/OM values approach zero. This can make it an unsuitable function to use for low OC/OM soils, and makes this function especially sensitive to starting values. For example, with PTF-30 this error seemed to be

associated with the addition of the natural log transformed depth term; functions which were identical except for that term (i.e., PTF-24, PTF-25, PTF-26, PTF-27, PTF-36, PTF-37, PTF-40) had been successfully recalibrated.

Lastly, dataset size affects the ability of NLS to reach convergence. In addition to the three datasets presented in this study, multiple smaller datasets were generated and used to test PTFs. These datasets were based on carbon content or depth interval, with the largest being 116 horizons. However, NLS frequently terminated and generated error messages when these datasets were used; when convergence was successful and new coefficients generated, the accuracy was low and uncertainty high. It is therefore recommended that the use of small datasets be avoided.

2.4.4. Selection of PTFs for Recalibration

When approaching the task of selecting a PTF for use in predicting missing data, the first step is to consider which other variables are available. Choosing a smaller number of variables which results in a larger dataset has shown to be an advantage. In this study, recalibrated PTFs showed the best performance and the lowest uncertainty when recalibrated on the largest dataset, which was the Ontario (All Variable) dataset. Similarity of the data will also affect the performance and uncertainty estimations; the Ontario data was from a much smaller area, with predominantly agricultural soils, while the BC datasets were drawn from an area more than three hundred times greater in size and from a wide variety of ecological zones.

Choosing which variables to include also may depend on the region for which the PTF is being developed. In this study, OC and OM were shown to be the most important variables to include, which is consistent with other studies which developed PTFs for similar regions. For example, PTFs-36, 37 and 38 (Hossain et al., 2015) were developed for Arctic and sub-Arctic soils in Canada; PTFs-7, 8, 9, and 10 (Alexander, 1989) were developed for soils in Alaska. All these PTFs only contained OC as an input variable and showed high performance both before and after recalibration. OM has shown to be influential on bulk density since early PTFs (Saini, 1966), and for soils found in very different environments, such as the tropical rainforest (Gharahi Ghehi et al., 2012) The many PTFs that use only OC or OM attest to the strong inverse relationship between OC/OM and bulk density. Further, Minasny and Hartemink (2011) recommended that the

PTF with the fewest parameters should be chosen, and functions with simple model forms and only one input variable, such as those in groups A, B, C, D, and E, performed well across datasets in this study. Other variables may be important for different soil forming conditions, for example, texture and pH variables were included in PTFs developed for Brazil (Barros et al., 2015; Bernoux et al., 1998; Beutler et al., 2017; Pereira et al., 2016; Tomasella and Hodnett, 1998). While these PTFs performed well in their region of development, inclusion of texture and pH as input variables did not improve their performance in this study, even after recalibration.

Using the uncertainty metrics of MPI and PICP graphs can be useful when choosing between model forms. In this study, the coefficients were recalibrated for every PTF, and the MPI results of model forms in Groups A to D are shown in Figure 2.5. From this information, PTFs from Group B would have the lowest uncertainty when recalibrated on the Ontario dataset, while PTFs from Group D have the lowest uncertainty when recalibrated on both BC datasets. Finally, accuracy metrics such as RMSE and CCC can be used to identify suitable PTFs; PTFs from Groups B, C, D and E showed strong performance across datasets. Linear model forms (Group A) performed slightly less well, and model forms which included other soil attributes such as texture, pH or coarse fragment showed results that varied significantly. Overall, PTFs which contained only OM/OC as an input, and which showed a non-linear relationship between OC/OM and bulk density, performed the best and had the lowest uncertainty of the PTFs which were tested.

2.5. Conclusion

Using bulk density as the example target variable, 73 PTFs were recalibrated using an NLS approach, on multiple case study datasets. It was demonstrated that NLS, with few exceptions, improves the accuracy of the PTF, potentially significantly. The performance of the recalibrated functions depended on multiple factors – the model form; the type of input variable; and the dataset which was used. PTFs which have fewer input variables, and minimal transformations of those variables (i.e., log transformations, multiple exponents, or trigonometric functions) are more easily recalibrated with NLS. The best performing recalibrated PTFs were those with only OC or OM as input variables; and which were recalibrated on the largest dataset. When recalibrating PTFs in the future, these factors must be taken into account when choosing a PTF model form

for recalibration. The most suitable model for use would be one with few terms, using starting values that are carefully chosen.

Uncertainty estimates generated through quantile regression demonstrate that the uncertainty of a PTF is dataset dependent; PTFs recalibrated on the largest dataset all had low uncertainty, while those recalibrated on the smallest dataset had the highest uncertainty. This is a potential limitation to the use of NLS; it is not recommended to recalibrate PTFs on small datasets, as this will reduce the accuracy and increase the uncertainty of the resulting PTF.

Other studies have shown the usefulness of including environmental covariates to PTFs, such as climatic and topographic variables (Schillaci et al., 2021), and it could be beneficial for future studies to explore the inclusion of environmental variables when developing PTFs. However, when recalibrating existing PTFs, the choice of PTF may be limited by the variables available in the case study dataset, as was the case in this study; the legacy datasets used did not include environmental variables.

The framework presented here will allow users to select a PTF based on the input variables that are available, after consideration of the region or soil conditions for which the PTF was developed. With a lack of regional PTFs available for many parts of the world, especially Canada, recalibration of existing PTFs through NLS provides an accessible framework to generate a PTF suitable for the dataset for which it is being used. The coupling of recalibration with quantile regression to produce uncertainty estimates in the same process can then allow the user to select a PTF with minimal uncertainty, and to present the uncertainty estimates along with the accuracy metrics and final recalibrated PTF.

2.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

2.7. Acknowledgements

The authors acknowledge the financial support that was provided by the Forest Innovation Program of the Canadian Wood Fibre Centre, Natural Resources Canada; the support from the British Columbia Ministry of Water and Land Resources Stewardship; and the British Columbia Ministry of Forests, Lands, Natural Resource Operations and Rural Development. A. Arbor was funded by a Canadian Graduate Scholarship from the Natural Sciences and Engineering Research Council of Canada.

For the Ottawa study area, the authors acknowledge the financial support that was provided by the Ontario Ministry of Agriculture, Food and Rural Affairs through the Growing Forward 2 program. Contributors to the field and lab activities include: Jim Warren, Adam Gillespie, Stephanie Vickers, Mackenzie Clarke, and Veronika Wright.

2.8. Tables and Figures

Table 2.1. Summary statistics of all datasets: Ontario (All Variables), BC (All Variables), BC (Carbon and Bulk Density)

Dataset	Attribute	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Ontario (All Variables) (n = 3242)	Bulk Density (g/cm^3)	0.30	0.99	1.18	1.17	1.35	1.89
	Organic Carbon (%)	0.03	0.84	1.89	2.29	2.87	16.90
	Depth (Midpoint, cm)	2.00	11.00	14.00	21.45	33.00	112.50
	pH (H_2O)	3.82	6.05	6.86	6.73	7.54	8.34
	Coarse Fragment (%)	0.00	0.00	1.00	4.28	5.00	65.00
	Silt (%)	0.00	24.90	35.80	34.96	45.30	86.40
	Sand (%)	0.90	20.95	41.30	42.24	60.90	99.20
	Clay (%)	0.0	10.90	18.60	22.80	31.20	83.20
BC (All Variables) (n = 396)	Bulk Density (g/cm^3)	0.42	1.07	1.32	1.33	1.56	2.66
	Organic Carbon (%)	0.05	0.37	1.01	1.92	2.22	17.06
	Depth (Midpoint, cm)	1.10	10.00	28.50	40.86	71.50	142.00
	pH (H_2O)	4.40	5.40	5.80	5.94	6.30	8.90
	Coarse Fragment (%)	0.00	1.04	15.00	27.58	40.00	658.20
	Silt (%)	0.70	21.56	37.00	37.87	55.25	92.76
	Sand (%)	2.00	26.18	43.70	48.54	71.00	97.57
	Clay (%)	0.72	6.18	12.08	13.77	19.00	72.00
BC (Carbon and Bulk Density) (n = 1199)	Bulk Density (g/cm^3)	0.35	1.08	1.29	1.33	1.55	2.66
	Organic Carbon (%)	0.00	0.46	1.18	2.19	2.57	17.06

Table 2.2. List of studies which have compared the performance of PTFs for soil bulk density.

Reference	Region/Soil	No. PTFs Tested	Evaluation Metrics
Abdelbaki, 2018	USA	48	ME, RMSE, EF (modelling efficiency)
Boschi et al., 2018	Brazil	25	MAE, ME, RMSE, R^2, REC
Casanova et al., 2016	Chile	10	R^2
De Vos et al., 2005	Flanders, Belgium	12	MPE, SDPE, RMSPE, R^2
Han et al., 2012	China	19	MPE, RMSPE, SDPE, R^2
Kaur et al., 2002	India	12	ME, RMSPD
Nanko et al., 2014	Japanese volcanic soils	29	MPE, SDPE, RMSPE, R^2, % of prediction data within +/- 10 and 20% in relative error
Nasta et al., 2020	Italy	63	RMSE, R^2
Reidy et al., 2016	Ireland	12	MPE, SDPE, RMSPE, R^2
Sevastas et al., 2018	Greece	56	ME, MAE, SDPE, RMSE
Taulya et al., 2005	Uganda	8	ME, RMSPD, R^2
Vasiliniuc and Patriche, 2015	Romania	22	MPE, RMSPE, SDPE, R^2
Yi et al., 2016	Qinhai Province, China	14	ME, SDE, RMSE, R^2

Legend: Mean prediction error (MPE); standard deviation of the prediction error (SDPE); root mean square prediction error (RMSPE); coefficient of determination (R^2); regression error curve (REC); mean absolute error (MAE); root mean square prediction difference (RMSPD); and mean error (ME).

Table 2.3. Equation-based PTFs from the literature that were tested and recalibrated in this study.

PTF #	Study	PTF	Soils/Region	Units	Evaluation	Sample size
1	Abdelbaki, 2018	$BD = 1.449e^{-0.03*OC}$	USA	BD in Mg/m^3, OC in %	R^2 0.680 RMSE 0.13 EF 0.59 ME 0.0	45,195
2	Akpa, 2016	BD = 1.177 + 0.00263*Sand - 0.0439*log(Silt) + 0.00208*Silt	Nigeria (All data)	BD in Mg/m^3, Sand and Silt in %	MAE 0.140 RMSE 0.179 R^2 0.109 CCC 0.185	1161
3	Akpa. 2016	BD = 1.172 + 0.0025*Sand - 0.0341*log(Silt) + 0.000877*Silt	Nigeria (Topsoil)	BD in Mg/m^3, Sand and Silt in %	MAE 0.116 RMSE 0.152 R^2 0.161 CCC 0.255	1161
4	Akpa, 2016	BD = 1.512–0.00322*Clay-0.0865*log(Silt)	Nigeria (Subsoil)	BD in Mg/m^3, Sand and Silt in %	MAE 0.151 RMSE 0.189 R^2 0.139 CCC 0.25	1161
5	Alexander, 1980	$BD = 1.66 – 0.308*OC^{0.5}$	Upland soil, California USA	BD in g/cm^3, OC in %	R^2 0.462 SE 0.19 g/cm^3	386 total
6	Alexander, 1980	$BD = 1.72 – 0.294 \times OC^{0.5}$	Alluvial soil, California, USA	BD in g/cm^3, OC in %	R^2 0.332 SE 0.16 g/cm^3	335 total
7	Alexander, 1989	$BD = 1.83*\exp(-0.121*OC^{0.5})$	Alaska, USA	BD in Mg/m^3, OC in %	R^2 0.81	55
8	Alexander, 1989	$BD = 2.24*\exp(-0.120*OC^{0.5})$	Loamy sands; Alaska, USA	BD in Mg/m^3, OC in %	R^2 0.85	55
9	Alexander, 1989	$BD = 1.86*\exp(-0.120*OC^{0.5})$	Sandy loams; Alaska, USA	BD in Mg/m^3, OC in %	R^2 0.85	55
10	Alexander, 1989	$BD = 1.73*\exp(-0.120*OC^{0.5})$	Silt loams; Alaska, USA	BD in Mg/m^3, OC in %	R^2 0.85	55
11	Barros et al., 2015	BD = 1.495 – 0.011*Clay – 0.079*pH	Amazonia, Brazil	BD in kg/dm^3, pH in H_2O 1:1	R^2 0.73 AIC -250.29	140 for training, 125 for testing

PTF #	Study	PTF	Soils/Region	Units	Evaluation	Sample size
12	Bernoux et al., 1998	BD = 1.52 – 0.0038*Clay – 0.050*OC – 0.045*pH + 0.0010*Sand	Overall PTF; Brazil soils	BD in Mg/m^3, Clay, Sand and OC in %, pH in H_2O 1:1	R^2 0.56 SE 0.09	323
13	Bernoux et al., 1998	BD for sandy soils = 0.0181*Sand – 0.08*OC	Sandy soils; Brazil	BD in Mg/m^3, OC and Sand in %	R^2 0.66	14
14	Beutler et al., 2017	$BD = [1.6179 – 0.0180*(Clay+1)^{0.46} – 0.0398*OC^{0.55}]^{1.33}$	Organic soil with clay, Brazil	BD in Mg/dm^3, clay in g/kg, OC in g/kg	ME -0.04 RMSE 0.22 R^2 0.47	280
15	Beutler et al., 2017	$BD = [4.0899 – 2.3978*OC^{0.06}]^{3.85}$	Organic soil Brazil	BD in Mg/dm^3, OC in g/kg	ME -0.10 RMSE 0.26 R^2 0.37	280
16	Botula et al., 2015	BD = 1.64581 – 0.00362*Clay – 0.0016*Sand – 0.0158*OC	Congo, Africa	BD in Mg/m^3, OC in %, Clay and Sand in %	R^2 0.244 MAD 0.110 Mg/m^3 RMSD 0.137 Mg/m^3	196 training; 72 test
17	Brahim et al., 2012	BD = 1.65 – 0.117*OC – 0.0042*Clay – 0.0036*Coarse Sand + 0.031*pH	All horizons; Tunisia, Africa	BD in Mg/m^3, OC, Clay, Coarse Sand in %, pH in H_2O 1:1	R^2 0.55 SE 0.14	348
18	Brahim et al., 2012	BD = 0.9 – 0.08*OC + 0.007*FineSand + 0.007*FineSilt + 0.05*pH	Superficial horizons (<40 cm depth). Tunisia, Africa	BD in Mg/m^3, OC, Fine Silt, Fine Sand in %, pH in H_2O 1:1	R^2 0.58 SE 0.14	286
19	Curtis and Post, 1964	$Y = 2.09963 – 0.00064X_1 – 0.22302X_2$ $Y = \log(BD*100)$ $X_1 = \log(\%LOI)$ $X_2 = X_1^2$	Vermont, USA	OM % BD g/cm^3	R 0.96 SE 0.054	78
20	De Vos et al., 2005	$BD = 1.775 – 0.173(OM)^{0.5}$	Belgium	BD in Mg/m^3, LOI in %	R^2 0.57 MPE -0.004 Mg/m^3 RMSPE 0.16 Mg/m^3	1614
21	Drew, 1973	BD = 1/(0.6268 + 0.0361*(OM))	Minnesota	OM in %, BD in g/cc	R^2 0.842	80

PTF #	Study	PTF	Soils/Region	Units	Evaluation	Sample size
22	Federer, 1983	$\ln(BD) = -2.314 - 1.0788*\ln(OM) - 0.1132*(\ln OM)^2$	New England, USA	BD in Mg/m^3, OM in g/g	Not given; in previous study	130
23	Gosselink et al. 1984	BD = 0.026/OC *100	Louisiana marsh soils	BD in g/ml, OC in %	R^2 0.93	288
24	Grigal et al., 1989	BD = 0.073 + 2.369*exp(-0.073*OM)	Forest floor. North-central USA	BD in Mg/m^3, LOI in %	R^2 0.75	812
25	Grigal et al., 1989	BD = 0.669 + 0.941*exp(-0.240*OM)	Surface mineral soil (0-25 cm). North-central USA	BD in Mg/m^3, LOI in %	R^2 0.95	800
26	Grigal et al., 1989	BD = 0.043X + 4.258*exp(-0.047*OM)	Peat, where X = 0 for surface peat (0-25 cm) and X = 1 for 25-175cm depth peat	BD in Mg/m^3, LOI in %	R^2 0.89	232
27	Grigal et al., 1989	BD = 0.075 + 1.301*exp(-0.060*OM)	All data. North-central USA.	BD in Mg/m^3, LOI in %	R^2 0.93	1612
28	Han et al., 2012	$\ln(BD) = 0.5379 - 0.0653*OM^{0.5}$	China	Db in g/cm^3, OM in g/g	MPE 0.0 RMSPE 0.13 SDPE 0.13 R^2 67.1	Training 75% Data; Validation 25%
29	Heinonen, 1977	BD = 1.42 – 0.0016*Clay + 0.0021*Silt	Finland Glaciofluvial soils at 30-50 cm depth	BD in g/cm^3, Silt and Clay in %	R 0.79	20
30	Hollis et al., 2012	BD = 1.5868 – (0.4682*exp(0.0578*OC)) – (0.07778*ln(horizon mid-point)	Volcanic materials. Europe	BD in g/cm^3, OC in %	RMSE 0.17 g/cm^3, model efficiency 0.44	34 training; 3 validation
31	Hollis et al., 2012	BD = 0.80806 + (0.823844*exp(-0.27993* OC) + (0.0014065*Sand) – (0.0010299*Clay)	Cultivated topsoils (mineral A horizons). Europe	BD in g/cm^3, Sand, Clay and OC in % mass	RMSE 0.13 g/cm^3, model efficiency 0.62	333 training; 126 validation
32	Hollis et al., 2012	BD = 1.1257 – (0.1140245*Ln(OC)) + (0.0555*Ln(Horizon mid-point)) + (0.002248*Sand)	Compact subsoils. Europe	BD in g/cm^3, OC in % mass	RMSE 0.14 g/cm^3, model efficiency 0.40	64 training; 55 validation

PTF #	Study	PTF	Soils/Region	Units	Evaluation	Sample size
33	Hollis et al., 2012	BD = 0.69794 + (0.750636* exp(−0.230355*OC) + (0.0008687*Sand) – (0.0005164*Clay)	All other mineral horizons. Europe	BD in g/cm^3, Sand, Silt, Clay and OC in % mass	RMSE 0.15 g/cm^3, model efficiency 0.63	925 training; 604 validation
34	Hollis et al., 2012	BD = 1.4903 – 0.33293* ln(OC)	All organic horizons. Europe	BD in g/cm^3, Sand, Silt, Clay and OC in % mass	RMSE 0.10 g/cm^3, model efficiency 0.68	67 training; 24 validation
35	Honeysett and Ratkowsky, 1989	$1/\rho_b = 0.564 + 0.0556$ I BD = 1/(0.564 + 0.556*OM)	Forest soils, Tasmania	BD in g/cm^3, I is % oven-dried weight of OC		136 Four sampling depths: 0-5 cm; 0-10 cm; 25-30 cm; 75-80 cm
36	Hossain et al., 2015	BD = 0.701 + 0.952*exp(-0.29*OC)	Mineral soil. Arctic and sub-Arctic of Canada	BD in g/cm^3, SOC in %	R^2 0.99	702
37	Hossain et al., 2015	BD = 0.074 + 2.632*exp(-0.076*OC)	Organic soil. Arctic and sub-Arctic of Canada	BD in g/cm^3, OC in %	R^2 0.93	674
38	Hossain et al., 2015	BD = 0.071 + 1.322*exp(-0.071*OC)	Combined mineral and organic soils. Arctic and sub-Arctic of Canada	BD in g/cm^3, OC in %	R^2 0.984	1376
39	Huntington et al., 1989	ln (BD) = -2.39 – 1.316*ln(OM) – 0.167*(ln(OM))^2	New Hampshire	BD in Mg/m^3, OC in %	R^2 0.75	60 profiles, with depth intervals and horizons
40	Huntington et al., 1989	ln (BD) = 0.263 – 0.147*ln(OC) – 0.103*(ln(OC))^2	New Hampshire	BD in Mg/m^3, OC in %	R^2 0.72	60 profiles, with depth intervals and horizons
41	Jeffrey 1970	BD = 1.482 – 0.6786*log(OM)	Compilation – Ohio, England, Europe, Australia	BD in g/ml, LOI in %	r = -0.9045	80
42	Katterer et al., 2006	BD = 1.6384 – 0.0945*OC	Sweden (Topsoil model 3)	Topsoil is 0-25 cm depth; subsoil is 25-100 cm depth. BD in g/cm^3; OC in %	R^2 0.51 RMSE 0.13	337

PTF #	Study	PTF	Soils/Region	Units	Evaluation	Sample size
43	Katterer et al., 2006	BD = 1.6444 – 0.1195*OC	Sweden (Subsoil model 3)	BD in g/cm^3; OC in %	R^2 0.45 RMSE 0.16	1283
44	Katterer et al., 2006	BD = 1.6693 – 0.1168*OC + 0.0391*(OC*Clay)	Sweden (Topsoil model 4)	BD in g/cm^3; OC and Clay in %	R^2 0.47 RMSE 0.14	337
45	Katterer et al., 2006	BD = 1.5994 + 0.1111*Clay – 0.0787*OC - 0.0857*(OC*Clay)	Sweden (Subsoil model 4)	BD in g/cm^3; OC and Clay in %	R^2 0.46 RMSE 0.16	1283
46	Katterer et al., 2006	BD = 1.5815 + 0.1171*(Sand+Gravel) – 0.1078*OC	Sweden (Topsoil model 5)	BD in g/cm^3; OC, Sand, CF in %	R^2 0.53 RMSE 0.13	337
47	Katterer et al., 2006	BD = 1.6270 + 0.0965*(Sand+Gravel) – 0.1608*OC	Sweden (Subsoil model 5)	BD in g/cm^3; OC, Sand, CF in %	R^2 0.46 RMSE 0.16	1283
48	Katterer et al., 2006	BD = 1.6333 – 0.1233*OC + 0.0433*(OC*Sand)	Sweden (Topsoil model 6)	BD in g/cm^3; OC, Sand in %	R^2 0.53 RMSE 0.12	337
49	Katterer et al., 2006	BD = 1.6552 -0.0561*Sand – 0.1525*OC -0.0533*(OC*Clay) + 0.1160*(OC*Sand)	Sweden (Subsoil model 6)	BD in g/cm^3; OC, Sand, Clay in %	R^2 0.49 RMSE 0.16	1283
50	Kaur et al., 2002	BD = exp(0.313 – 0.191*OC + 0.02102*Clay – 0.000476*(Clay^2) – 0.00432*Silt	India (agricultural, pine forest, oak forest, barren)	OC in %(g/g), Sand, Silt, Clay in % (g/g), BD in g/cm^3	Adj. R^2 0.62 SE 0.25 ME 0.0 RMSPD 0.15 g/cm^3	112 training, 112 validation
51	Kobal et al., 2011	BD = 1.4842 – 0.1424*OC (for OC < 3.6%) BD = 1.1253 – 0.0452*OC (for OC >3.6%)	Slovenia Forest mineral soils	OC g/kg BD in g/cm^3	R^2 0.7958 SE 0.1257	109
52	Makovníková et al., 2017	BD = 3.1482 – 0.0118*Clay – 0.017*Sand – 0.0152*Silt	Slovakia, 0-10 cm depth	BD in g/cm^3, OC in %, Silt (0.001-0.05mm), Sand (0.05-2mm) and Clay (<0.01mm) in %	SD 0.084 g/cm^3 CV R^2 0.27	262
53	Makovníková et al., 2017	BD = 2.662 – 0.0076*Clay – 0.0102*Silt – 0.0108*Sand – 0.0855*OC	Slovakia, 0-10 cm depth	BD in g/cm^3, OC in %, Silt (0.001-0.05mm), Sand (0.05-2mm) and Clay (<0.01mm) in %	SD 0.110 g/cm^3 CV R^2 0.46	262

PTF #	Study	PTF	Soils/Region	Units	Evaluation	Sample size
54	Manrique and Jones, 1991	$BD = 1.510 – 0.113*OC$	USA (incl. Hawaii and Puerto Rico); other countries.	BD in Mg/m^3, OC in %	R^2 0.36	19 651
55	Manrique and Jones, 1991	$BD = 1.660 – 0.318*OC^{0.5}$	USA (incl. Hawaii and Puerto Rico); other countries.	BD in Mg/m^3, OC in %	R^2 0.41	19 651
56	Nanko et al., 2014	$BD = 1 / (0.882 + 0.133*OC)$	Japan. Forest soil affected by volcanic ash.	BD in g/cm^3, OC in %	MPE -0.0021 SDPE 0.138 RMSPE 0.138 R^2 67.3	Training: 3513 Validation: 279
57	Pereira et al., 2016	$BD = 1.326 + 0.315*\sin(1.045 – 0.001*Clay – 0.052*OC) + 0.0003*Clay*\sin(\sin(2.561 + 1.287*pH – 0.0006*Clay)) – 0.134*\sin(\sin(2.561 + 1.287*pH – 0.006*Clay))$	Brazilian Amazon forest soils.	BD in g/cm^3, others in g/kg	MSE 0.015 RMSE 0.123	Training dataset: 654 Validation dataset: 230
58	Premrov et al., 2018	$BD^{0.71} = 13.801 – 13.446*OM^{0.02}$	Ireland	BD, OM in $[g/cm^3]^{0.71}$	RMSPE 0.126 AIC_{corr} -56	111 training 28 testing
59	Prévost, 2004	$\ln(BD) = -1.81 – 0.892*\ln(OM) - 0.092*(\ln(OM))^2$	Forest soils. "Logarithmic approach"	BD in Mg/m^3, OM in g/g	R^2 0.767	318, 32, 89, 86 Two sites, two treatments
60	Qiao et al., 2019	$BD = 1.68 + 0.001*Depth – 2.249*Clay^{-1} – 0.089*Depth^{-1}$	China; Loess Plateau deep layers (50-200m)	BD in g/cm^3, OC in g/kg, Silt, Clay in %	R^2 0.356 RMSE 0.079 ME -0.008	Training: 427; Validation: 107
61	Ruehlmann and Körschens, 2009	$BD = a*\exp(b*OC)$ Original: $BD = (2.684 – 140.943*b)\exp(-b*OC)$	Arable, Reclaimed, Wetland. (b = 0.008 in original)	BD in Mg/m^3, OC in g/kg	R^2 0.872 RMSE 0.215	59 datasets used to create dataset used; n ranged from 3 to 193. Summary stats for each source available in paper.
62	Ruehlmann and Körschens, 2009	$BD = a*\exp(b*OC)$ Original: $BD = (2.684 – 140.943*b)\exp(-b*OC)$	Proctor. (b = 0.006 in original)	BD in Mg/m^3, OC in g/kg	R^2 0.872 RMSE 0.215	59 datasets used to create dataset used; n ranged from 3 to 193. Summary stats for each source available in paper.

PTF #	Study	PTF	Soils/Region	Units	Evaluation	Sample size
63	Ruehlmann and Körschens, 2009	BD = a*exp(b*OC) Original: BD = (2.684 – 140.943*b)exp(-b*OC)	Volcanic. (b = 0.010 in original)	BD in Mg/m^3, OC in g/kg	R^2 0.872 RMSE 0.215	59 datasets used to create dataset used; n ranged from 3 to 193. Summary stats for each source available in paper.
64	Saini, 1966	BD = 1.62 – 0.06*OM	Humic-gley	BD in g/cm^3, OM in %	r -0.858	30
65	Saini, 1966	BD = 1.53 – 0.05*OM	Imperfectly drained	BD in g/cm^3, OM in %	r -0.805	40
66	Saini, 1966	BD = 1.52 – 0.06*OM	Well drained	BD in g/cm^3, OM in %	r -0.633	40
67	Sevastas et al., 2018	BD = 2.268 – 0.179*ln(Sand) – 0.345*ln(OC)	River basin in Northern Greece	BD in g/cm^3; OC in %; use 2 as OC:OM conversion factor	ME -0.00188 MAE 0.0823 SDPE 0.1215 RMSE 0.1195	30
68	Sevastas et al., 2018	BD = 2.039 – 0.563*OC + 0.103*OC^2	River basin in Northern Greece	BD in g/cm^3, OC in %; use 2 as OC:OM conversion factor	ME 0.00016 MAE 0.09636 SDPE 0.12662 RMSE 0.12449	30
69	Song et al., 2005	BD = 1.3565*$e^{-0.0046*OC}$	China, uncultivated soils	BD in g/cm^3, OC in g/kg	R^2 0.726	3645
70	Song et al., 2005	BD = 1.3770*$e^{-0.0048*OC}$	China, cultivated soils	BD in g/cm^3, OC in g/kg	R^2 0.787	4765
71	Tamminen and Starr, 1994	BD = 1.565 – 0.2298*$(OM)^{0.5}$	Finland	BD in kg/dm^3, OM in %	R^2 0.61	75 samples for 0-5 cm; 60 for 30-35 cm; 23 for 60-65 cm
72	Tomasella and Hodnett, 1998	BD = 1.578 – 0.054*OC – 0.006*Silt – 0.004*Clay	Brazil	BD in g/cm^3, OC, Silt, Clay in %	R^2 0.774	396
73	Yanti et al., 2021	BD = 1.2684 + 0.0011*Depth – 0.1774*OC	Indonesia	BD in g/cm^3, OC in %, Depth in cm	R^2 0.425 ME 0.048 RMSE 0.120	45

Legend: Mean prediction error (MPE); standard deviation of the prediction error (SDPE); root mean square prediction error (RMSPE); coefficient of determination (R^2); regression error curve (REC); mean absolute error (MAE); root mean square prediction difference (RMSPD); and mean error (ME).

Table 2.4. List of model groups by type and reference. Model groups are based on model form and input variables used. Where soil organic carbon (OC) is indicated as an input variable, soil organic matter (OM) may also be substituted. If the PTF used OM, this is indicated in the reference column with (OM).

Model Group	Model Type	References (PTF #)
A (Linear)	$BD = a - b*OC$	Katterer et al. 2006 (42), (43) Kobal et al., 2011 (51) Manrique and Jones 1991 (54) Saini 1966 (OM) (64), (65), (66)
B (Radical root)	$BD = a - b*OC^{0.5}$	Alexander 1980 (5), (6) De Vos 2005 (20) (OM) Manrique and Jones 1991 (55) Tamminen and Starr 1994 (71) (OM)
C (Reciprocal)	$BD = 1 / (a + b*OC)$	Honeysett and Ratkowsky 1989 (35) (OM) Nanko et al. 2014 (56) Drew, 1973 (21) (OM)
D (Multiple ln terms)	$\ln(BD) = a + b*\ln(OC) + c*[\ln(OC)]^2$	Curtis and Post 1964 (19) Federer 1983 (22) Huntington et al. 1989 (39 OM), (40 OC) Prévost 2004 (59)
E (Natural exponent)	$BD = a + b*\exp(c*OC)$	Grigal et al. 1989 (24), (25), (26), (27) (OM) Hossain et al. 2015 (36), (37), (38)
	$BD = a*\exp(b*OC)$	Adelbaki 2018 (1) Grigal et al., 1989 (26) (OM) Song et al., 2005 (69), (70) Ruehlmann and Körschens, 2009 (61), (62), (63)
	$BD = a*\exp(b*OC^{0.5})$	Alexander 1989 (7), (8), (9), (10)
	$BD = \exp(a - b*OM^{0.5})$	Han et al 2012 (28)
F (With only OM/OC)	$BD = [a - b*OC^{0.06}]^{3.85}$	Beutler et al., 2017 (15)
	$BD = a/OC*100$	Gosselink et al., 1984 (23)
	$BD = a - b*\ln(OC)$	Hollis et al., 2012 (34)
	$BD = a - b*\log(OM)$	Jeffrey 1970 (41) (OM)
	$BD^{0.71} = (a - b*OM^{0.02})$	Premrov et al., 2018 (58)
	$BD = a - b*OC + c*OC^2$	Sevastas et al., 2018 (68)

G **(With OC/OM, other terms)**	BD = a – b*Clay – c*OC – d*pH + e*Sand	Bernoux et al., 1998 (12)
	BD = a*Sand – b*OC	Bernoux et al., 1998 (13)
	BD = [a – b*(Clay+1)$^{0.46}$ – c*OC$^{0.55}$]$^{1.33}$	Beutler et al., 2017 (14)
	BD = a – b*Clay – c*Sand – d*OC	Botula et al., 2015 (16)
	BD = a – b*OC – c*Clay – d*Sand + e*pH	Brahim et al., 2012 (17)
	BD = a – b*OC + c*Sand + d*Silt + e*pH	Brahim et al., 2012 (18)
	BD = a + (b*exp^(c*OC)) + d*Sand – e*Clay	Hollis et al., 2012 (31), (33)
	BD = a – (b*ln(OC)) + (c*ln(Depth)) + (d*Sand)	Hollis et al., 2012 (32)
	BD = a – b*OC + c*(OC*Clay)	Katterer et al., 2006 (44)
	BD = a + b*Clay – c*OC – d*(OC*Clay)	Katterer et al., 2006 (45)
	BD = a + b*(Sand + CF) – c*OC	Katterer et al., 2006 (46), (47)
	BD = a – b*OC + c*(OC*Sand)	Katterer et al., 2006 (48)
	BD = a – b*Sand – c*OC – d*(OC*Clay) + e*(OC*Sand)	Katterer et al., 2006 (49)
	BD = exp(a – b*OC + c*Clay – d*(Clay^2) – e*Silt)	Kaur et al., 2002 (50)
	BD = a – b*Clay – c*Silt – d*Sand – e*OC	Makovníková et al., 2017 (53)
	BD = a – b*ln(Sand) – c*ln(OC)	Sevastas et al., 2018 (67)
	BD = a – b*OC – c*Silt - d*Clay	Tomasella and Hodnett, 1998 (72)
	BD = a + b*Depth – c*OC	Yanti et al., 2021 (73)
H **With only texture, pH or depth**	BD = a – b*Clay – c*pH	Barros et al 2015 (12)
	BD = a – b*Clay + c*Silt	Heinonen 1977 (29)
	BD = a – b*Clay – c*Sand – d*Silt	Makovníková et al., 2017 (52)
	BD = a + b*Depth - c*(1/Clay) – d*(1/Depth)	Qiao et al., 2019 (60)
X **Could not NLS**	BD = a – (b*exp(c*OC)) – (d*log(Depth))	Hollis et al 2012 (30)
	BD = (a + b*sin(c - d*OC) + (e*Clay)*sin(sin(f + g*pH - h*Clay)) - j*sin(sin(k + m*pH - n*Clay)))	Pereira et al., 2016 (57)

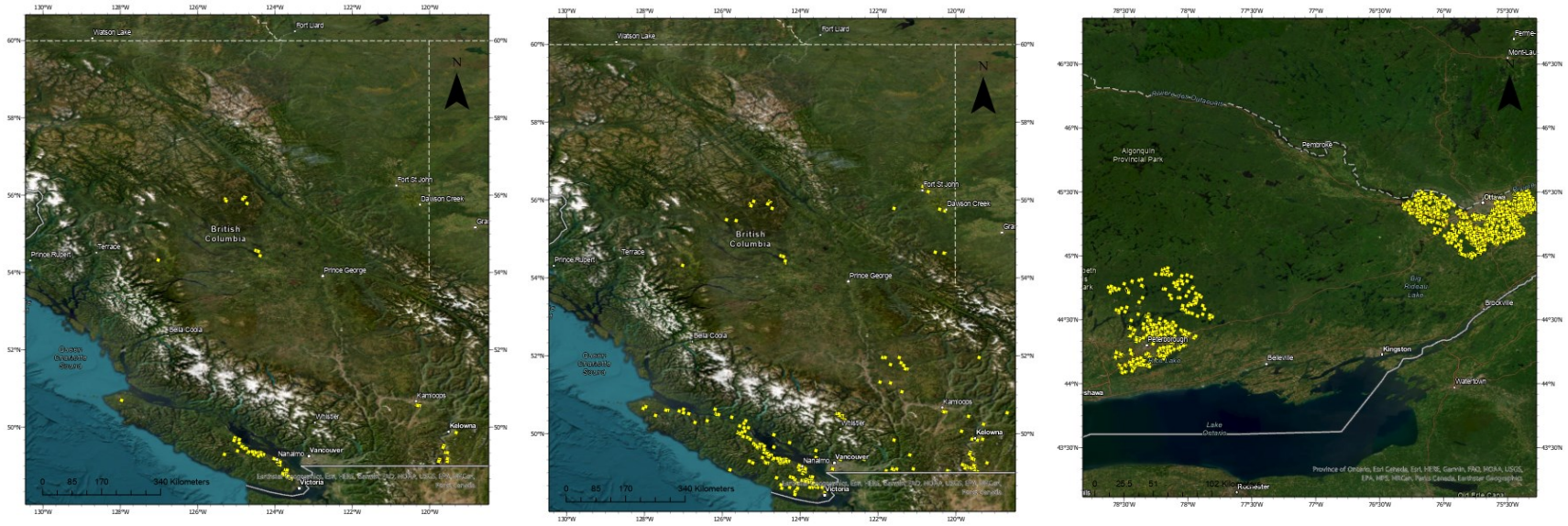


Figure 2.1. Sample site locations for the BC (All Variables) dataset (left); the BC (Carbon and Bulk Density) dataset (middle); and the Ontario (All Variables) dataset (right).

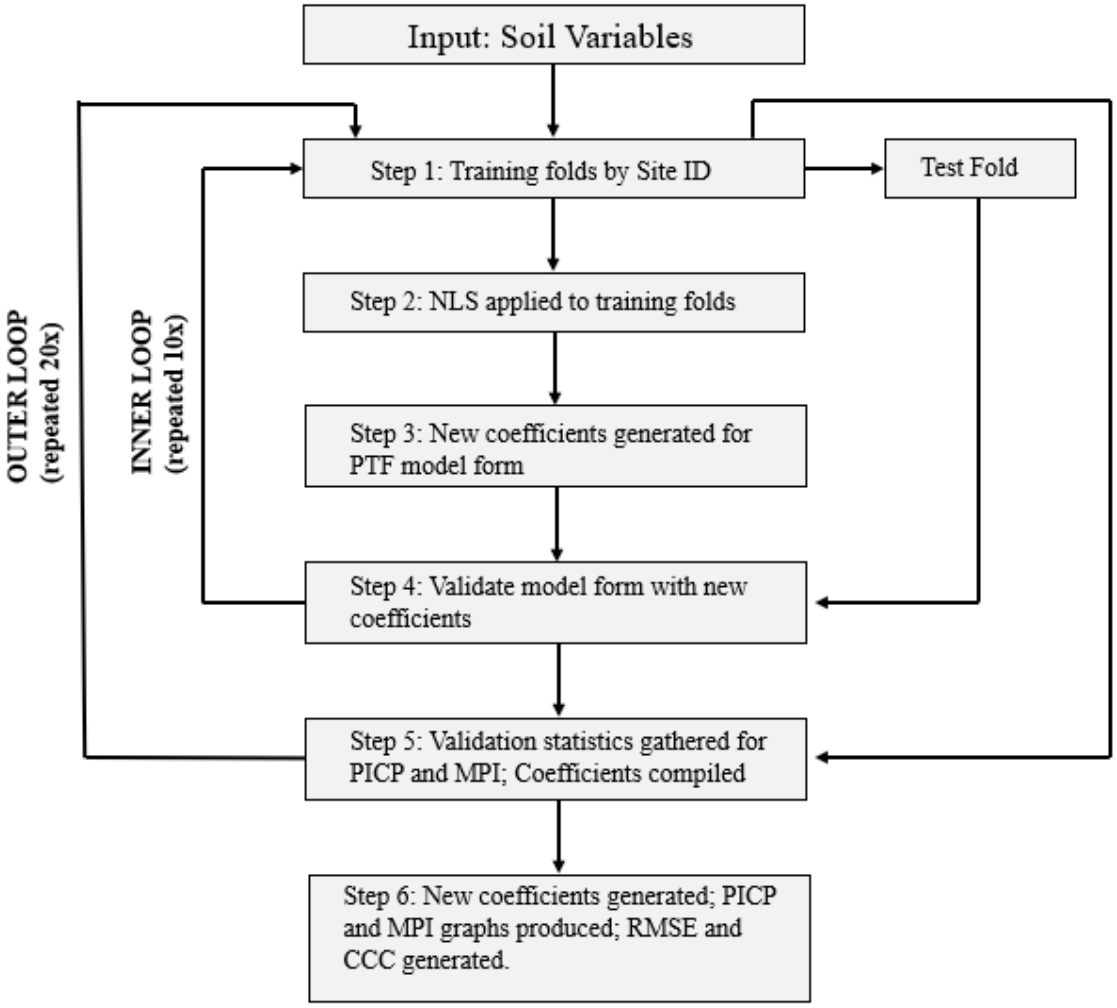


Figure 2.2. Schematic of the nested cross-validation which produces new coefficients for existing model forms through non-linear least squares; procedure also generates uncertainty estimates through the quantile regression.

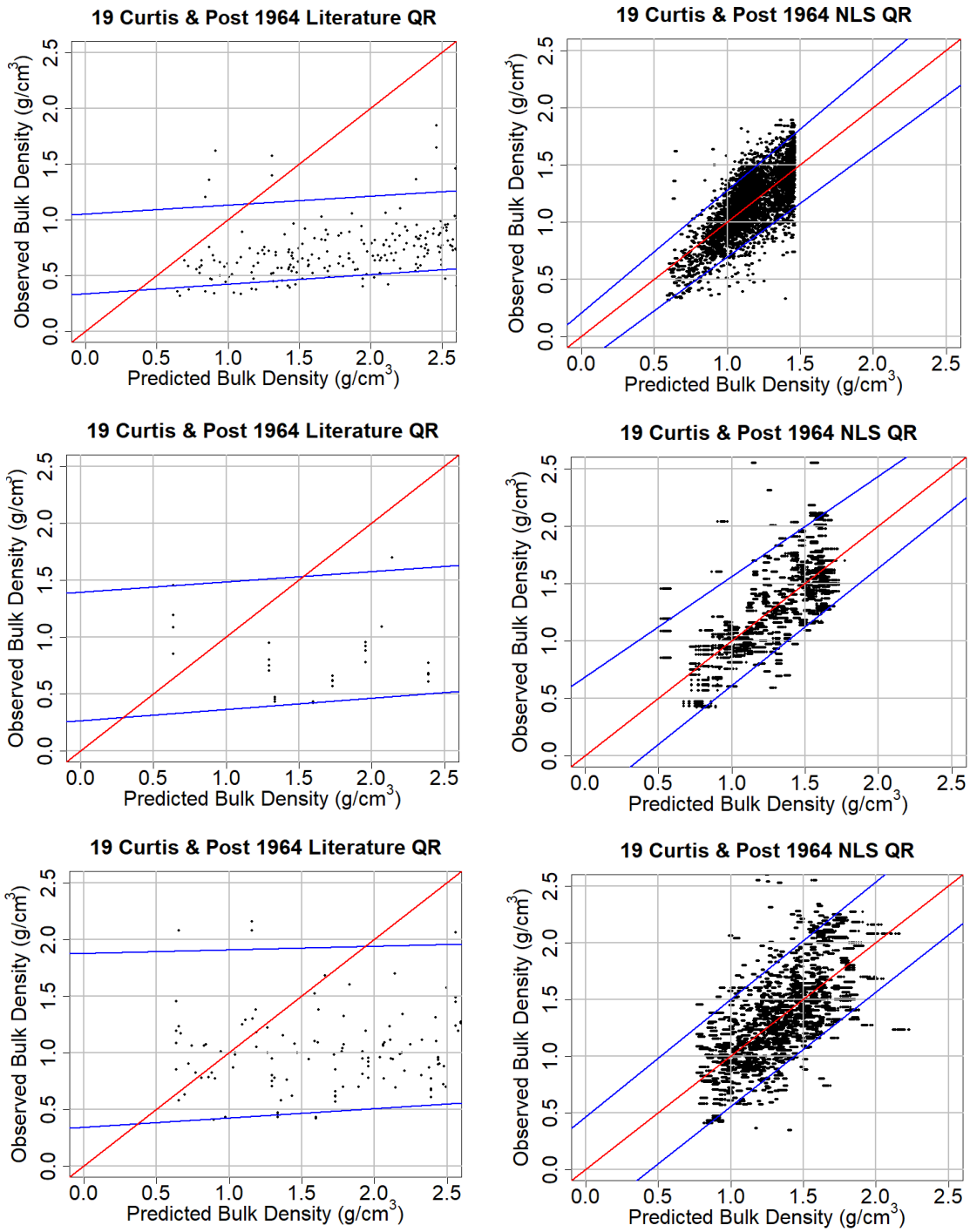


Figure 2.3. **Plots of observed versus predicted BD values for Group D (multiple natural log terms) functions. For each dataset used in this study, an example PTF was chosen for the model. This function was plotted using its literature coefficients, and then using its NLS-generated coefficients. Blue lines represent the 5th and 95th percentiles; red line is the 1:1 line. Ontario (All Variables) results are top left and right; BC (All Variables) are middle left and right; BC (Carbon and Bulk Density) results are bottom left and right.**

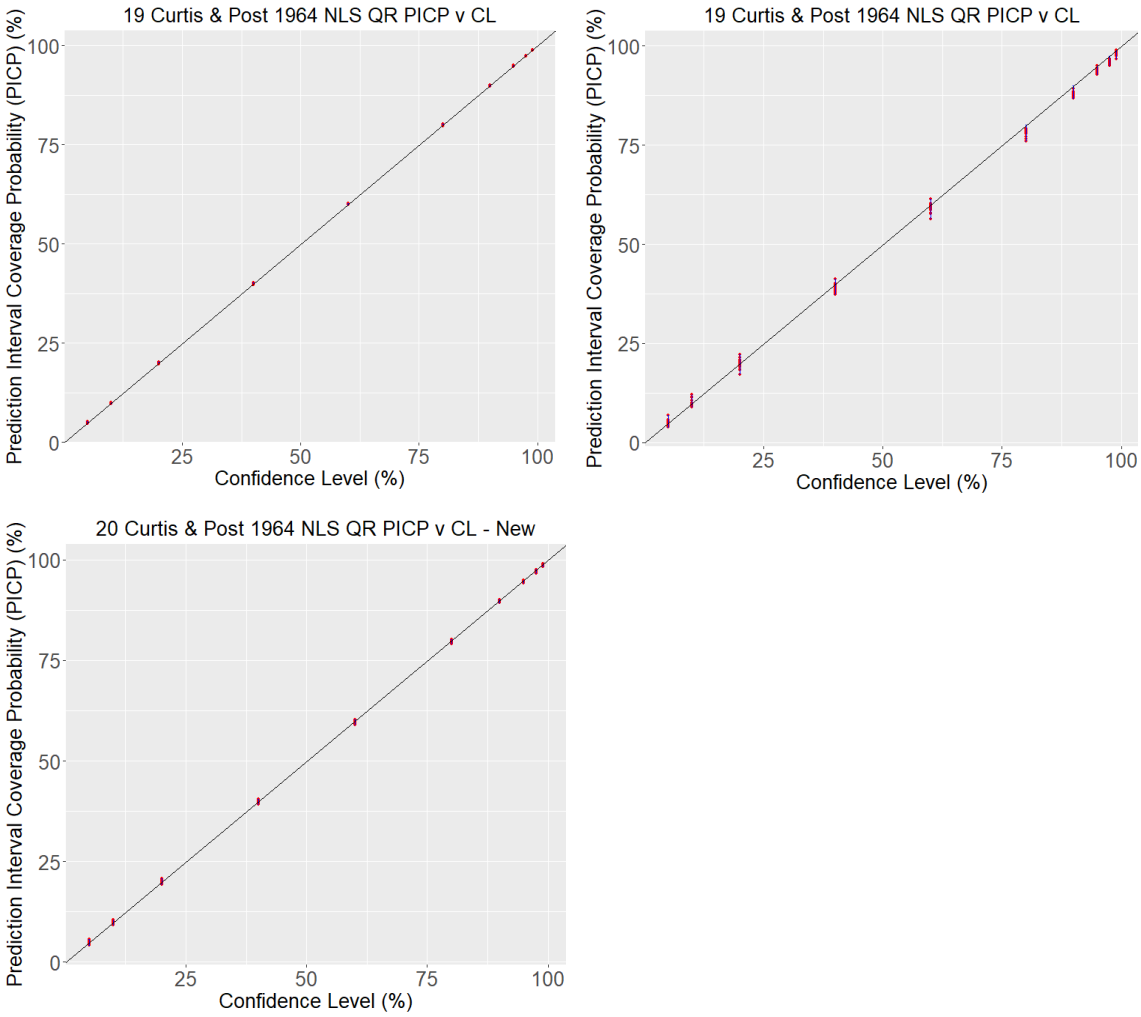


Figure 2.4. **PICP vs CL Graphs for each dataset. For Model Group D (functions with multiple natural log terms), a representative PTF was chosen, and the PICP vs CL graph for that PTF is shown for each dataset. Ontario (All Variables) is shown top left; BC (All Variables) is shown top right; and BC (Carbon and Bulk Density) is shown bottom left.**

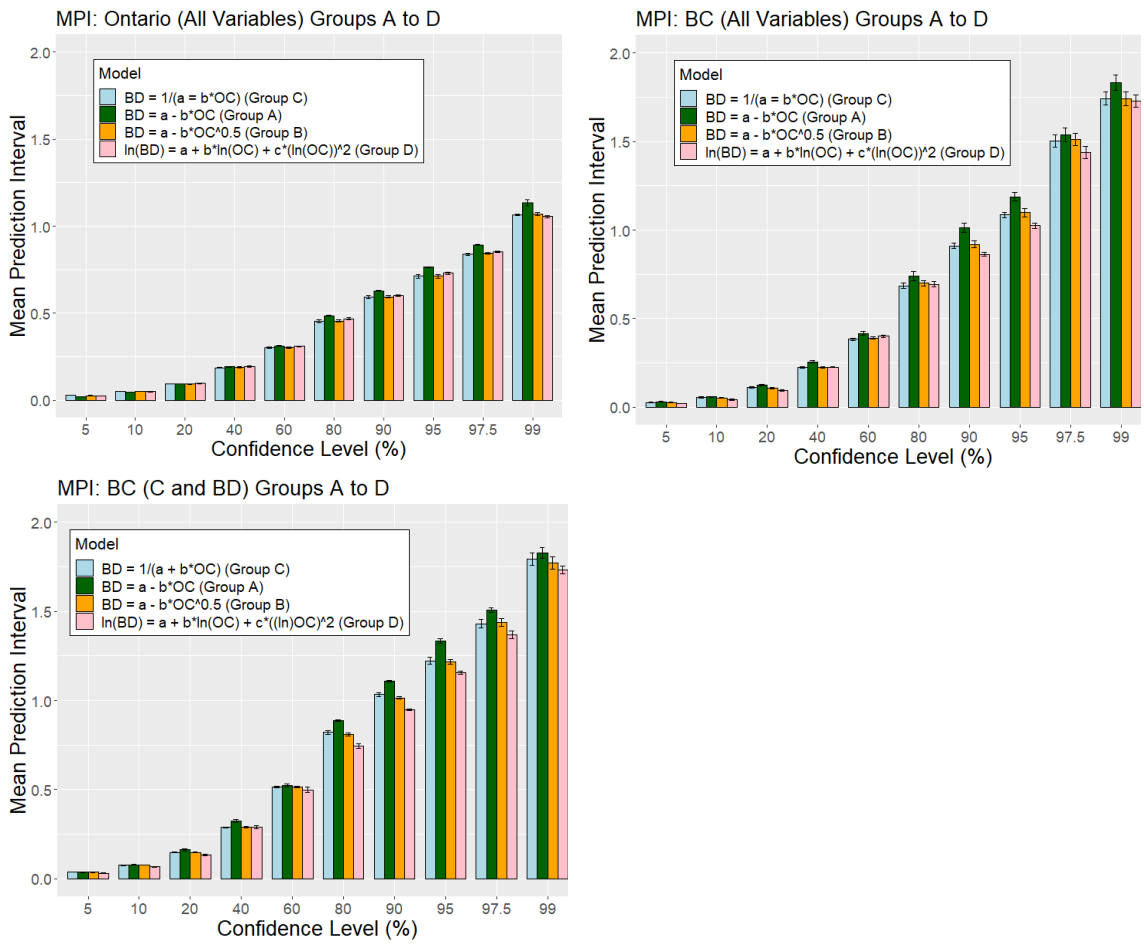


Figure 2.5. **MPI Graphs for Model Groups A to D, for each dataset. For each MPI graph, the MPI values for each recalibrated model in Groups A to D are shown. Groups A to D have PTFs whose model form is identical within the group, and the recalibrated PTFs for each model group have the same MPI values. Results for Ontario (All Variables) is shown top left; BC (All Variables) is shown top right; and BC (Carbon and Bulk Density) is shown bottom left.**

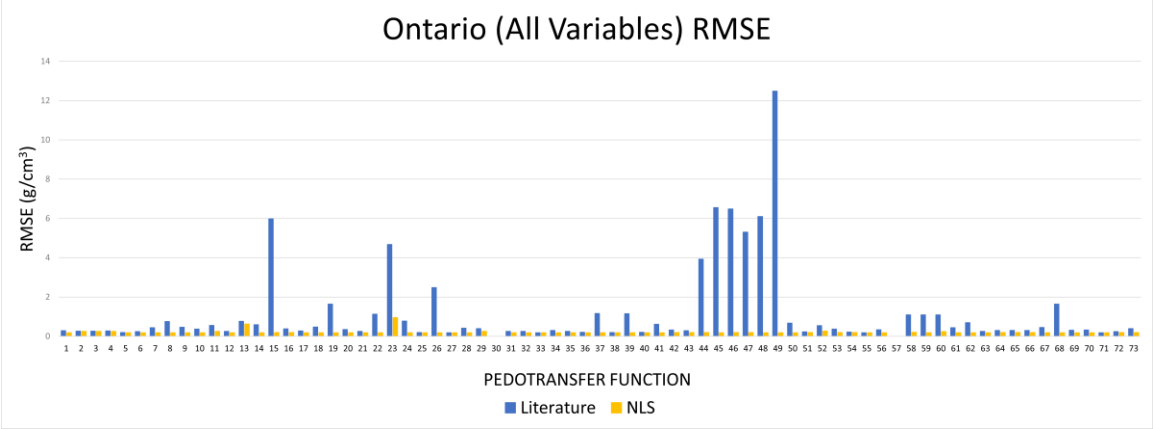


Figure 2.6. RMSE values produced by PTFs before and after recalibration on the Ontario (All Variables) dataset.

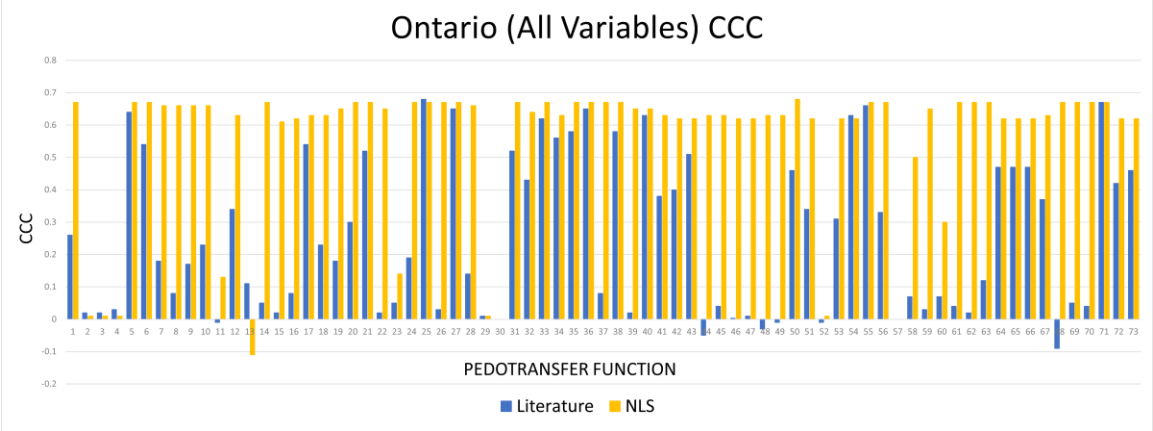


Figure 2.7. CCC values produced by PTFs before and after recalibration on the Ontario (All Variables) dataset.

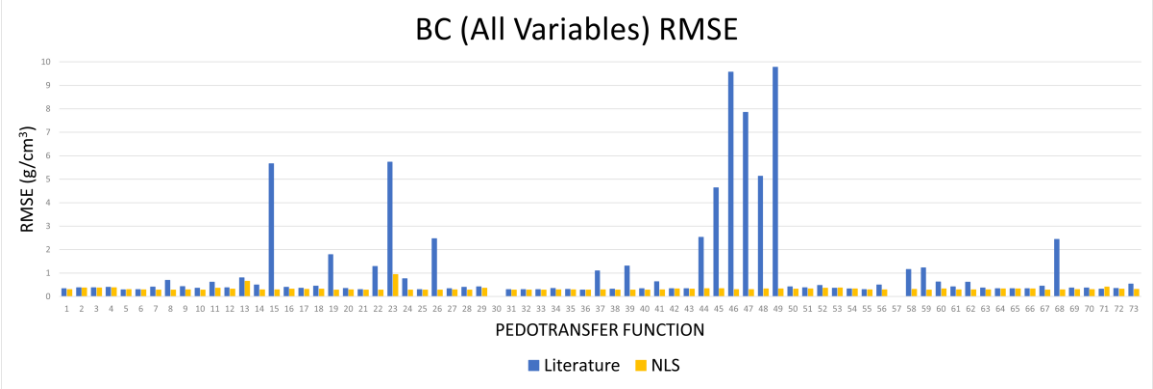


Figure 2.8. RMSE values produced by PTFs before and after recalibration on the BC (All Variables) dataset.

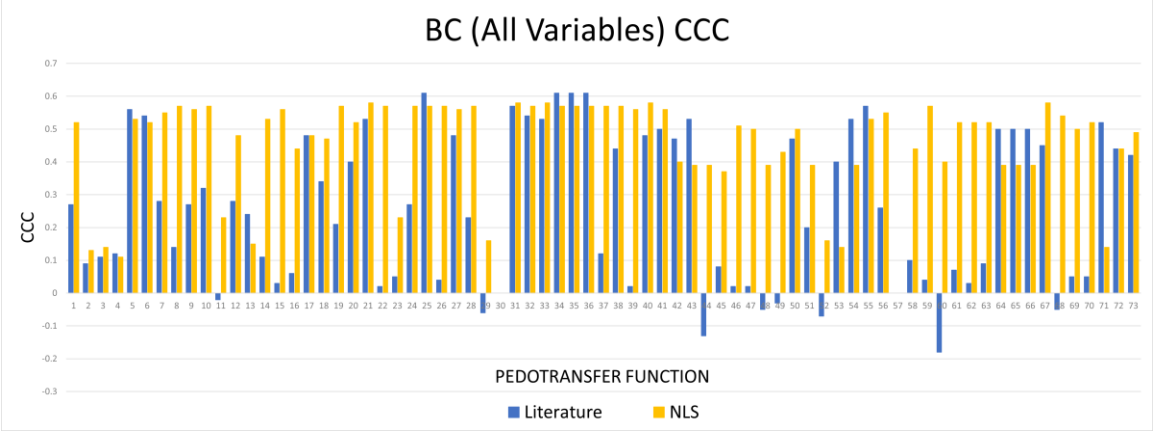


Figure 2.9. CCC values produced by PTFs before and after recalibration on the BC (All Variables) dataset.

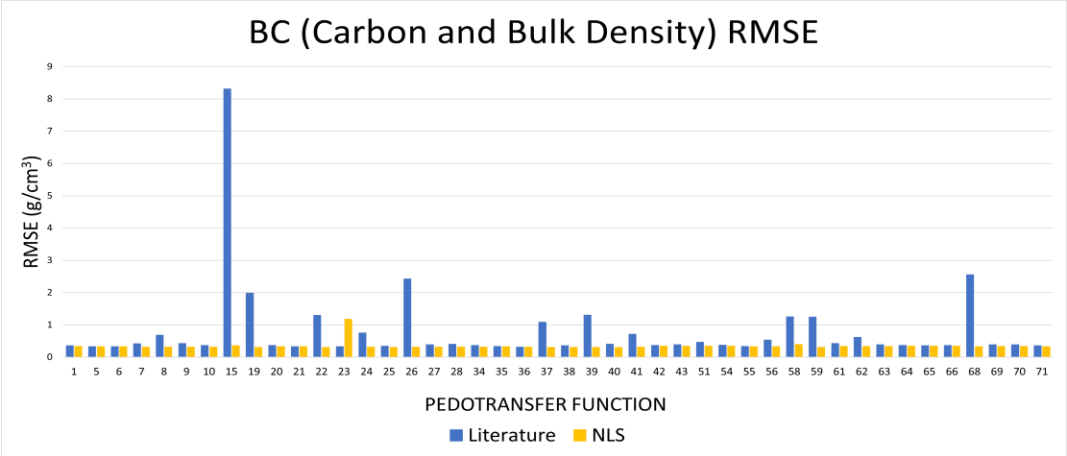


Figure 2.10. RMSE values produced by PTFs before and after recalibration on the BC (Carbon and Bulk Density) dataset.

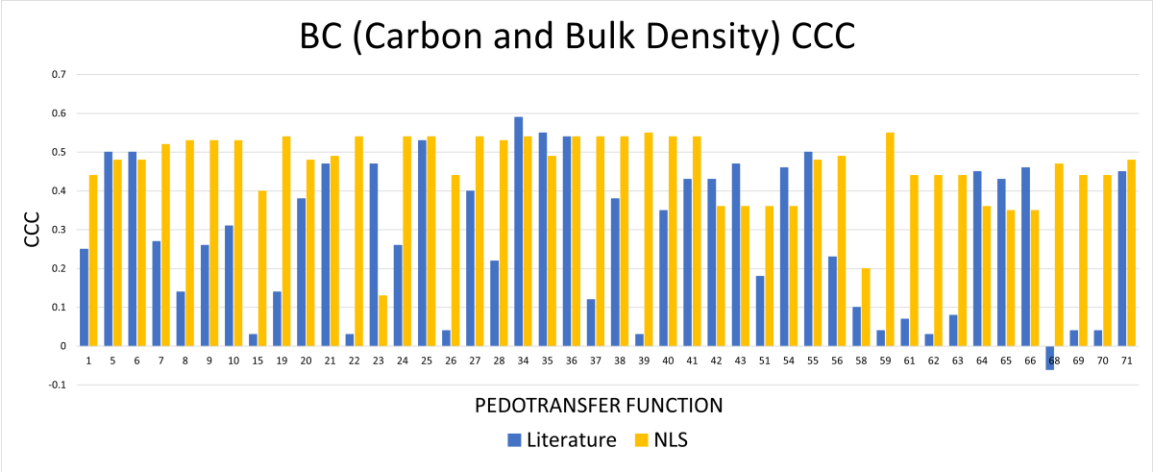


Figure 2.11. CCC values produced by PTFs before and after recalibration on the BC (Carbon and Bulk Density) dataset.

2.9. References

Adams, W.A. 1973. The effect of organic matter on the bulk and true densities of some uncultivated podzolic soils. Journal of Soil Science, 24: 10-17.

Abdelbaki, A.M. 2018. Evaluation of pedotransfer functions for predicting soil bulk density for U.S. soils. Ain Shams Engineering Journal, 9: 1611-1619.

Akpa, S.I.C., Ugbaje, S.U., Bishop, T.F.A. and Odeh, I.O.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: 644-658.

Alaboz, P., Demir, S., Dengiz, O. 2020. Assessment of various pedotransfer functions for the prediction of the dry bulk density of cultivated soils in a semiarid environment. Communications in Soil Science and Plant Analysis, 52(7): 724-742

Alexander, E.B. 1980. Bulk densities of California soils in relation to other soil properties. Soil Sci. Soc. Am. J. 44: 689-692.

Alexander, E.B. 1989. Bulk density equations for southern Alaska soils. Canadian Journal of Soil Science, 69(1): 177-180.

Al-Qinna, M.I. and Jaber, S.M. 2013. Predicting soil bulk density using advanced pedotransfer functions in an arid environment. Transactions of the ASABE, 56(3): 963-976.

Arrouays, D., Leenaars, J.G.B., Richer-de-Forges, A.C., Adhikari, K., Ballabio, C., Greve, M., Grundy, M., Guerrero, E., Hempel, J., Hengl, T., Heuvelink, G., Batjes, N., Carvalho, E., Hartemink, A., Hewitt, A., Hong, S.-Y., Krasilnikov, P., Lagacherie, P., Lelyk, G., Libohova, Z., Lilly, A., McBratney, A., McKenzie, N., Vasquez, G.M., Mulder, V.L., Minasny, B., Montanarella, L., Odeh, I., Padarian, J., Poggio, L., Roudier, P., Saby, N., Savin, I., Searle, R., Solbovoy, V., Thompson, J., Smith, S., Sulaeman, Y., Vintila, R., Rossel, R.V., Wilson, P., Zhang, G.-L., Swerts, M., Oorts, K., Karklins, A., Feng, L., Navarro, A.R.I., Levin, A., Laktionova, T., Dell'Acqua, M., Suvannang, N., Ruam, W., Prasad, J., Patil, N., Husnjak, S., Pásztor, L., Okx, J., Hallett, S., Keay, C., Farewell, T., Lilja, H., Juilleret, J., Marx, S., Takata, Y., Kazuyuki, Y., Mansuy, N., Panagos, P., Van Liedkerke, M., Skalsky, R., Sobocka, J., Kobza, J., Eftekhari, K., Alavipanah, S.K., Moussadek, R., Badraoui, M., Da Silva, M., Paterson, G., Gonçalves, M.d.C., Theocharopoulos, S., Yemefack, M., Tedou, S., Vrscaj, B., Grob, U., Kozák, J., Boruvka, L., Dobos, E., Taboada, M., Moretti, L., and Rodriquez, D. 2017. Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. GeoResJ, 14: 1-19.

Ball, D.F. 1964. Loss-on-ignition as an estimate of organic matter and organic carbon in non-calcareous soils. Journal of Soil Science, 15(1): 84-92.

Baritz, R., Seufert, G., Montanarella, L. and Van Ranst, E. 2010. Carbon concentrations and stocks in forest soils of Europe. Forest Ecology and Management, 260: 262-277.

Barros, H.S. and Fearnside, P.M. 2015. Pedo-transfer functions for estimating soil bulk density in central Amazonia. R. Bras. Ci. Solo, 39: 397-407.

Bates, D.M. and Watts, D.G. Nonlinear regression analysis and its applications. John Wiley and Sons, Inc. USA, 1999.

Batjes, N.H., Van Engelen, V.W.P., Kauffmann, J.H., and Oldeman, L.R. 1994. Development of soil databases for global environmental modelling. In Trans. 15th World Congress of Soil Science (Acapulco, Mexico, 10-17 July, 1994), 6: 40-57.

Benites, V.M., Machado, P.L.O.A., Fidalgo, E.C.C., Coelho, M.R. and Madari, B.E. 2007. Pedotransfer functions for estimating soil bulk density from existing soil survey reports in Brazil. Geoderma, 139: 90-97.

Bernoux, M., Arrouays, D., Cerri, C., Volkoff, B. and Jolivet, C. 1998. Bulk Densities of Brazilian Amazon soils related to other soil properties. Soil Sci. Soc. Am. J., 62: 743-749.

Beutler, S.J., Pereira, M.G., Tassinari, W.d.S., Menezes, M.D., Valladares, G.S., Anjos, L.H.C. 2017. Bulk density prediction for histosols and soil horizons with high organic matter content. Rev. Bras. Cienc. Solo., 41: e0160158

Blake, G.R. Ch. 30, Bulk Density. In Methods of Soil Analysis, Part 1. 1965. C.A. Black, Editor-in-Chief, and D.D. Evans [and Others] Associate Editors; R.C. Dinauer, Managing Editor. American Society of Agronomy. 374-390

Boschi, R.S., Bocca, F.F., Lopes-Assad, M.L.R.C., Assad, E.D. 2018. How accurate are pedotransfer functions for bulk density for Brazilian soils? Scientia Agricola, 75(1): 70-78.

Botula, Y-D., Nemes, A., Van Ranst, E., Mafuka, P., De Pue, J., and Cornelis, W.M. 2015. Hierarchical pedotransfer functions to predict bulk density of highly weathered soils in Central Africa. Soil Sci. Soc. Am. J. 79: 476-486.

Bouma, J. 1989. Using soil survey data for quantitative land evaluation. Advances in Soil Science, 9: 177-213.

Brahim, N., Bernoux, M. and Gallali, T. 2012. Pedotransfer functions to estimate soil bulk density for Northern Africa: Tunisia case. Journal of Arid Environments, 81: 77-83.

Cade, B.S. and Noon, B.R. 2003. A gentle introduction to quantile regression for ecologists. Front. Ecol. Environ., 1(8): 412-420.

Casanova, M., Tapia, E., Seguel, O., Salazar, O. 2016. Direct measurement and prediction of bulk density on alluvial soils of central Chile. Chilean Journal of Agricultural Research, 76(1): 105-113.

Chen, Y., Huang, Y. and Sun, W. 2017. Using organic matter and pH to estimate the bulk density of afforested/reforested soils in northwest and northeast China. Pedosphere, 27(5): 890-900.

Chen, S., Richer-de-Forges, A.C., Saby, N.P.A., Martin, M.P., Walter, C., Arrouays, D. 2018. Building a pedotransfer function for soil bulk density on regional dataset and testing its validity over a larger area. Geoderma, 312: 52-63.

Crins, W., Gray, P., Uhlig, P., Wester, M., 2009. The ecosystems of Ontario, Part 1: ecozones and ecoregions. Ont. Minist. Nat. Resour. Peterb. Ont. Inventory Monit. Assess. SIB TER IMA TR- 01 71pp.

Curtis, R.O. and Post, B.W. 1964. Estimating bulk density from organic-matter content in some Vermont forest soils. Soil Sci. Soc. Am. Proc. 28: 285-286.

De Vos, B., Van Meirvenne, M., Quataert, P., Deckers, J., and Muys, B. 2005. Predictive quality of pedotransfer functions for estimating bulk density of forest soils. Soil Sci. Soc. Am. J., 69: 500-510.

Dogulu, N., López López, P., Solomatine, D.P., Weerts, A.H., Shrestha, D.L. 2015. Estimation of predictive hydrologic uncertainty using the quantile regression and UNEEC methods and their comparison on contrasting catchments. Hydrology and Earth System Sciences, 19: 3181-3201.

Drew, L.A. 1973. Bulk density estimation based on organic matter content of some Minnesota soils. Minnesota Forestry Research Notes, 243. Sci. Jour. Ser. Paper No. 8333 of the University of Minnesota Agricultural Experimental Station.

Federer, C.A. 1983. Nitrogen mineralization and nitrification: depth variation in four New England forest soils. Soil Sci. Soc. Am. J. 47: 1008-1014.

Federer, C.A., Turcotte, D.E., and Smith, C.T. 1993. The organic fraction – bulk density relationship and the expression of nutrient content in forest soils. Can. J. For. Res. 23: 1026-1032.

Fox, J. and Weisberg, S. 2018. Nonlinear regression, nonlinear least squares, and nonlinear mixed models in R. Appendix to An R Companion to Applied Regression, 3rd ed.

Grigal, D.F., Brovold, S.L., Nord, W.S., and Ohmann, L.F. 1989. Bulk density of surface soils and peat in the north central United States. Can. J. Soil Sci., 69: 895-900

Gosselink, J.G., Hatton, R., and Hopkinson, C.S. 1984. Relationship of organic carbon and mineral content to bulk density in Louisiana marsh soils. Soil Science, 137(3): 177-180.

Han, G.-Z., Zhang, G.-L., Gong, Z.-T., and Wang, G.-F. 2012. Pedotransfer functions for estimating soil bulk density in China. Soil Science, 177(3): 158-164.

Heinonen, R. 1977. Towards “normal” soil bulk density. Soil Science Society of America Journal, 41(6): 1214-1215.

Hollis, J.M., Hannam, J., and Bellamy, P.H. 2012. Empirically-derived pedotransfer functions for predicting bulk density in European soils. European Journal of Soil Science, 63: 96-109.

Honeysett, J.L. and Ratkowsky, D.A. 1989. The use of ignition loss to estimate bulk density of forest soils. Journal of Soil Science, 40: 299-308.

Hossain, M.F., Chen, W., and Zhang, Y. 2015. Bulk density of mineral and organic soils in the Canada’s arctic and sub-arctic. Information Processing in Agriculture 2: 183-190.

Huntington, T.G., Johnson, C.E., Johnson, A.H., Siccama, T.G., and Ryan, D.F. 1989. Carbon, organic matter, and bulk density relationships in a forested spodosol. Soil Science, 148(5): 380-386.

Jeffrey, D.W. 1970. A note on the use of ignition loss as a means for the approximate estimation of soil bulk density. Journal of Ecology, 58(1): 297-299.

Kasraei, B., Heung, B., Saurette, D.D., Schmidt, M.G., Bulmer, C.E. and Bethel, W. 2021. Quantile regression as a generic approach for estimating uncertainty of digital soil maps produced from machine-learning. Environmental Modelling and Software, 144: 105139

Kätterer, T., Andrén, O., and Jansson, P.-E. 2006. Pedotransfer functions for estimating plant available water and bulk density in Swedish agricultural soils. Acta Agriculturae Scandinavica Section B-Soil and Plant Science, 56(4): 263-276.

Kaur, R., Kumar, S., and Gurung, H.P. 2002. A pedo-transfer function (PTF) for estimating soil bulk density from basic soil data and its comparison with existing PTFs. Aust. J. Soil Res., 40: 847-857.

Kobal, M., Urbančič, M., Potočič, N., De Vos, B., Simončič. 2011. Pedotransfer functions for bulk density estimation of forest soils. Journal of Forestry Soc. Croatia, 135: 19-27.

Koenker, R. and Bassett Jr., G. 1978. Regression Quantiles. Econometrica, 46(1): 33-50.

Kroetsch, D., and Wang, C. Particle size distribution. 2007. In: Carter, M.R. and Gregorich, E.G. (eds) Soil Sampling and Methods of Analysis, 2nd Ed. Canadian Society of Soil Science, Taylor & Francis, Florida, USA. pp. 713-725.

López López, P., Verkade, J.S., Weerts, A.H. and Solomatine, D.P. 2014. Alternative configurations of quantile regression for estimating predictive uncertainty in water level forecasts for the upper Severn River: a comparison. Hydrology and Earth System Sciences, 18: 311-3428.

Makovníková, J., Širáň, M., Houšková, B., Pálka, B. and Jones, A. 2017. Comparison of different models for predicting soil bulk density. Case study – Slovakian agricultural soils. International Agrophysics, 31: 491-498.

Malone, B.P., McBratney, A.B. and Minasny, B. 2011. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. Geoderma, 160: 614-626.

Manrique, L.A. and Jones, C.A. 1991. Bulk density of soils in relation to soil physical and chemical properties. Soil Sci. Soc. Am. J., 55: 476-481.

Marquardt, D.W. 1963. An algorithm for least-squares estimation of nonlinear parameters. J. Soc. Indust. Appl. Math., 11(2): 431-441.

McBratney, A.B., Minasny, B., Cattle, S.R., and Vervoort, R.W. 2002. From pedotransfer functions to soil inference systems. Geoderma, 109: 41-73.

McBratney, A.B., Minasny, B., and Tranter, G. 2011. Necessary meta-data for pedotransfer functions. Geoderma, 160: 627-629.

Minasny, B. and Hartemink, A.E. 2011. Predicting soil properties in the tropics. Earth-Science Reviews, 106: 52-62.

Muthusamy, M., Godiksen, P.N., and Madsen, H. 2016. Comparison of different configurations of quantile regression in estimating hydrological uncertainty. Procedia Engineering, 154: 513-520.

Nanko, K., Ugawa, S., Hashimoto, S., Imaya, A., Kob ayashi, M., Sakai, H., Ishizuka, S., Miura, S., Tanaka, N., Takahashi, M., Kaneko, S. 2014. A pedotransfer function for estimating bulk density of forest soil in Japan affected by volcanic ash. Geoderma, 213: 36-45.

Nasta, P., Palladino, M., Sica, B., Pizzolante, A., Trifuoggi, M., Toscanesi, M., Giarra, A., D'Auria, J., Nicodemo, F., Mazzitelli, C., Lazzaro, U., Di Fiore, P., Romano, N. 2020. Evaluating pedotransfer functions for predicting soil bulk density using hierarchical mapping information in Campania, Italy. Geoderma Regional, 21: e00267.

Nemes, A. 2015. Why do they keep rejecting my manuscript – do's and don'ts and new horizons in pedotransfer studies. Agrokémia és Talajtan, 64(2): 361-371.

Pereira, O.J.R., Montes, C.R., Lucas, Y. and Melfi, A.J. 2016. Evaluation of pedotransfer equations to predict deep soil carbon stock in tropical podzols compared to other soils of the Brazilian Amazon forest. In: Digital Soil Morphometrics, Progress in Soil Science, 331-349. A.E. Hartemink and B. Minasny (eds). Springer International Publishing, Switzerland.

Périé, C. and Ouimet, R. 2008. Organic carbon, organic matter, and bulk density relationships in boreal forest soils. Can. J. Soil Sci., 88: 315-325.

Premrov, A., Cummins, T., Byrne, K.A. 2018. Bulk-density modelling using optimal power-transformation of physical and chemical soil parameters. Geoderma, 314: 205-220.

Prévost, M. 2004. Predicting soil properties from organic matter content following mechanical site preparation of forest soils. Soil Sci. Soc. Am. K, 68: 943-949.

Qiao, J., Zhu, Y., Jia, X., Huang, L., and Shao, M. 2019. Development of pedotransfer functions for predicting the bulk density in the critical zone on the Loess Plateau, China. Journal of Soils and Sediments, 19: 366-372.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Rahmati, O., Choubin, B., Fathabadi, A., Coulon, F., Soltani, E., Shahabi, H., Mollaefar, E., Tiefenbacher, J., Cipullo, S., Bin Ahmad, B., and Bui, D.T. 2019. Predicting uncertainty of machine learning models for modelling nitrate pollution of groundwater using quantile regression and UNEEC methods. Science of the Total Environment, 688: 855-866.

Reidy, B., Simo, I., Sills, P. and Creamer, R.E. 2016. Pedotransfer functions for Irish soils – estimation of bulk density (ρb) per horizon type. SOIL, 2: 25-39.

Ritz, C., and Streibig, J.C. 2008. Nonlinear Regression with R. Giovanni Parmigiani (ed). Springer New York.

Ruehlmann, J. and Körschens, M. 2009. Calculating the effect of soil organic matter concentration on soil bulk density. Soil Sci. Soc. Am. J., 73: 876-885.

Saini, G.R. 1966. Organic matter as a measure of bulk density of soil. Nature, 210(5042): 1295-1296

Sevastas, S., Gasparatos, D., Botsis, D., Siarkos, I., Diamantaras, K.I. and Bilas, G. 2018. Predicting bulk density using pedotransfer functions for soils in the Upper Anthemountas basin, Greece. Geoderma Regional, 14: e00169.

Shreshta, D.L. and Solomatine, D.P. 2006. Machine learning approaches for estimation of prediction interval for the model output. Neural Networks, 19: 225-235.

Skjemstad, J.O. and Baldock, J.A. 2007. Chapter 21: Total and Organic Carbon. In: Soil Sampling and Methods of Analysis, (2nd ed). Carter, M.R. and Gregorich, E.G. (Eds.). CRC Press.

Soil Classification Working Group, 1998. The Canadian System of Soil Classification. NRC Research Press.

Song, G., Li, L., Pan, G. and Zhang, Q. 2005. Topsoil organic carbon storage of China and its loss by cultivation. Biogeochemistry, 74: 47-62.

Taulya, G., Tenywa, M.M., Majaliwa, M.J.G., Odong, T.L., Kaingo, J., and Kakone, A. 2005. Validation of pedotransfer functions for soil bulk density estimation on a Lake Victoria Basin soilscape. African Crop Science Conference Proceedings, 7: 1049-1056.

Tamminen, P. and Starr, M. 1994. Bulk density of forested mineral soils. Silva Fennica, 28(1): 53-60.

Teunissen, P.J.G. 1990. Nonlinear least squares. Manuscripta geodaetica, 15: 137-150.

Tomasella, J. and Hodnett, M.G. 1998. Estimating soil water retention characteristics from limited data in Brazilian Amazonia. Soil Science, 163(3): 190-202.

Tranter, G., Minasny, B., McBratney, A.B., Murphy, B., McKenzie, N.J., Grundy, M., and Brough, D. 2007. Building and testing conceptual and empirical models for predicting soil bulk density. Soil Use and Management, 23: 437-443.

Tremblay, S., Ouimet, R., and Houle, D. 2002. Prediction of organic carbon content in upland forest soils of Quebec, Canada. Can. J. For. Res., 32: 903-914.

Van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., Montzka, C., Nemes, A., Pachepsky, Y.A., Padarian, J., Schaap, M.G., Tóth, B., Verhoef, A., Vanderborght, J., van der Ploeg, M.J., Weihermüller, L., Zacharias, S., Zhang, Y., and Vereecken, H. 2017. Pedotransfer functions in earth system science: challenges and perspectives. AGU Reviews of Geophysics, 1199-1256.

Vasiliniuc, I. and Patriche, C.V. 2015. Validating soil bulk density pedotransfer functions using a Romanian dataset. Carpathian Journal of Earth and Environmental Sciences, 10(2): 225-236.

Walkley, A., and Black, I.A. 1934. An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. Soil Science, 37(1): 29-38.

Wang, D., and Anderson, D.W. 1998. Direct measurement of organic carbon content in soils by the Leco CR-12 carbon analyzer. Communications in Soil Science and Plant Analysis, 29: 15-21.

Yanti, E.D., Mulyono, A., Djuwansah, M.R., Narulita, I., Putra, R.D., Surinati, D. 2021. Development of pedotransfer functions for predicting soil bulk density: A case study in Indonesian small island. Journal of Water and Land Development, 51(X-XII): 181-187.

Yi, X.S., Li, G.S., and Yin, Y.Y. 2016. Pedotransfer functions for estimating soil bulk density: A case study in the Three-River Headwater region of Qinghai Province, China. Pedosphere, 26(3): 362-373.

2.10. Supplementary Tables

Table 2.5. Coefficients generated through NLS. PTFs grouped by model type. Where there is a mixture of PTFs which used OC and OM in a model group, the recalibrated coefficients are shown for the model form with both OC and OM versions. This is indicated for each coefficient in the model (ie the "a" coefficient is listed as both a_{OC} and a_{OM}).

Model Group	Model Form	References	Coeff. Ontario	Coeff. BC (All)	Coeff. BC (C and BD)
A (Linear)	BD = a – b*OC	Katterer et al. 2006 (42), (43) Kobal et al., 2011 (51) (OC) Manrique and Jones 1991 (54) (OC) Saini 1966 (OM) (64), (65), (66)	a_{OC}: 1.358 b_{OC}: 0.084 a_{OM} : 1.358 b_{OM} : 0.048	a_{OC}: 1.487 b_{OC}: 0.082 a_{OM}: 1.487 b_{OM}: 0.048	a_{OC}: 1.475 b_{OC}: 0.065 a_{OM}: 1.475 b_{OM}: 0.038
B (Radical root)	BD = a – b*$OC^{0.5}$ BD = a – b*$OM^{0.5}$	Alexander 1980 (5), (6) Manrique and Jones 1991 (55) (OC) De Vos 2005 (20) (OM) Tamminen and Starr 1994 (OM) (71)	a_{OC} : 1.590 b_{OC}: 0.307 a_{OM}: 1.59 b_{OM} : 0.234	a_{OC} : 1.732 b_{OC} : 0.341 a_{OM}: 1.732 b_{OM}: 0.259	a_{OC}: 1.695 b_{OC}: 0.287 a_{OM}:1.695 b_{OM}: 0.219
C (Reciprocal)	BD = 1 / (a + b*OM)	Nanko et al. 2014 (56) (OC) Honeysett and Ratkowsky 1989 (OM) (35) Drew, 1973 (21) (OM)	a_{OC}: 0.681 b_{OC}: 0.090 a_{OM} : 0.681 b_{OM}: 0.052	a_{OC}: 0.608 b_{OC}: 0.103 a_{OM}: 0.608 b_{OM}: 0.060	a_{OC}: 0.623 b_{OC}: 0.077 a_{OM}: 0.623 b_{OM}: 0.045
D (Multiple ln terms)	ln(BD) = a + b*ln(OM) + c*$[ln(OM)]^2$	Huntington et al. 1989 (39 OM), (40 OC) Curtis and Post 1964 (19) (OM)	a_{OC}: 0.255 b_{OC}: 0.148 c_{OC}: 0.045	a_{OC}: 0.312 b_{OC}: 0.176 c_{OC}: 0.038	a_{OC}: 0.273 b_{OC}: 0.157 c_{OC}: 0.008

		Federer 1983 (22) (OM) Prévost 2004 (59) (OM)	a_{OM}: 0.324 b_{OM}: 0.092 c_{OM}: 0.059	a_{OM}: 0.396 b_{OM}: 0.135 c_{OM}: 0.038	a_{OM}: 0.356 b_{OM}: 0.149 c_{OM}: 0.008
E (Natural exponent)	BD = a + b*exp(c*OC)	Hossain et al. 2015 (36), (37), (38)	a_{OC}: 0.466 b_{OC}: 0.995 c_{OC}: -0.178	a_{OC}: 0.769 b_{OC}: 0.925 c_{OC}: -0.403	a_{OC}: 1.023 b_{OC}: 0.809 c_{OC}: -0.910
		Grigal et al. 1989 (24), (25), (26), (27) (OM)	a_{OM}: 0.466 b_{OM}: 0.995 c_{OM}: -0.103	a_{OM}: 0.769 b_{OM}: 0.925 c_{OM}: -0.234	a_{OM}: 1.023 b_{OM}: 0.810 c_{OM}: -0.528
	BD = a*exp(b*OC)	Abdelbaki 2018 (1) Song et al., 2005 (69), (70) (OC) Grigal et al., 1989 (26) (OM)	a_{OC}: 1.427 b_{OC}: -0.097	a: 1.582 b: -0.109	a_{OC}: 1.537 b_{OC}: -0.075
			a_{OM}: 1.43 b_{OM}: -0.056	a_{OM}: 1.582 b_{OM}: -0.063	a_{OM}: 1.537 b_{OM}: -0.044
	BD = a*exp(b*$OC^{0.5}$)	Alexander 1989 (7), (8), (9), (10)	a: 1.68 b: -0.274	a: 1.882 b: -0.315	a: 1.826 b: -0.266
	BD = exp(a – b*$OM^{0.5}$)	Han et al 2012 (28)	a: 0.519 b: 0.209	a: 0.632 b: 0.240	a: 0.602 b: 0.203
	BD = a*exp(b*OC) Original, could not NLS: BD = (a – b*c)*exp(-c*OC)	Ruehlmann and Korschens 2009 (61), (62), (63)	a: 1.427 b: -0.097	a: 1.582 b: -1.088	a: 1.537 b: -0.075
F (With only OM/OC)	BD = (a – b*$OC^{0.06}$)$^{3.85}$	Beutler et al., 2017 (15)	a: 1.7302 b: 0.6727	a: 1.862 b: 0.787	a: 1.540 b: 0.460
	BD = a/OC*100	Gosselink et al., 1984 (23)	a: 35.88	a: 0.004	a: 0.002

	$BD = a - b*\log(OC)$	Hollis et al., 2012 (34)	a: 1.247 b: 0.187	a: 1.331 b: 0.1225	a: 1.357 b: 0.201
	$BD = a - b*\log(OM)$	Jeffrey, 1970 (41)	a: 1.348 b: 0.187	a: 1.453 b: 0.225	a: 1.466 b: 0.201
	$BD^{0.71} = a - b*OM^{0.02}$	Premrov et al., 2018 (58)	a: 7.622 b: 6.386	a: 8.770 b: 7.472	a: 4.442 b: 3.185
	$BD = a - b*OC + c*OC^2$	Sevastas et al., 2018 (68)	a: 1.441 b: 0.147 c: 0.006	a: 1.613 b: 0.205 c: 0.011	a: 1.586 b: 0.167 c: 0.009
G (With OC/OM, other terms)	$BD = a - b*Clay - c*OC - d*pH + e*Sand$	Bernoux et al., 1998 (12)	a: 1.12 b: 0.00007 c: 0.081 d: 0.033 e: 0.0002	a: 0.430 b: 0.008 c: 0.061 d: 0.110 e: 0.005	
	$BD = a*Sand - b*OC$	Bernoux et al., 1998 (13)	a: 0.018 b: 0.065	a: 0.021 b: 0.042	
	$BD = [a - b*(Clay+1)^{0.46} - c*OC^{0.55}]^{1.33}$	Beutler et al., 2017 (14)	a: 1.40344 b: -0.00078 c: 0.1993	a: 1.534 b: 0.020 c: 0.222	
	$BD = a - b*Clay - c*Sand - d*OC$	Botula et al., 2015 (16)	a: 1.3818 b: 0.0004 c: 0.0003 d: 0.0842	a: 1.166 b: -0.008 c: -0.004 d: 0.076	
	$BD = a - b*OC - c*Clay - d*Sand + e*pH$	Brahim et al., 2012 (17)	a: 1.12 b: 0.081 c: 0.00007 d: 0.00022 e: 0.033	a: 0.428 b: 0.061 c: -0.008 d:- 0.005 e: 0.110	
	$BD = a - b*OC + c*Sand + d*Silt + e*pH$	Brahim et al., 2012 (18)	a: 1.12 b: 0.081	a: 1.213 b: 0.061	

		c: 0.0001 d: 0.00008 e: 0.033	c: -0.0003 d: 0.008 e: 0.110
BD = a + (b*exp(c*OC)) + d*Sand – e*Clay	Hollis et al., 2012 (31), (33)	a: 0.508 b: 0.991 c: -0.179 d: -0.0005 e: 0.0008	a: 0.577 b: 0.880 c: -0.383 d: 0.003 e: -0.005
BD = a – (b*ln(OC)) + (c*ln(Depth)) + (d*Sand)	Hollis et al., 2012 (32)	a: 1.448 b: 0.220 c: 0.064 d: 0.0001	a: 1.132 b: 0.188 c: -0.032 d: -0.002
BD = a – b*OC + c*(OC*Clay)	Katterer et al., 2006 (44)	a: 1.365 b: 0.099 c: 0.0005	a: 1.486 b: 0.076 c: 0.0004
BD = a + b*Clay – c*OC – d*(OC*Clay)	Katterer et al., 2006 (45)	a: 1.411 b: 0.002 c: 0.110 d: 0.001	a: 1.489 b: 0.0003 c: 0.077 d: 0.0003
BD = a + b*(Sand + CF) – c*OC	Katterer et al., 2006 (46), (47)	a: 1.363 b: 0.0001 c: 0.084	a: 1.249 b: 0.003 c: 0.070
BD = a – b*OC + c*(OC*Sand)	Katterer et al., 2006 (48)	a: 1.363 b: 0.072 c: 0.0003	a: 1.481 b: 0.092 c: 0.003
BD = a – b*Sand – c*OC – d*(OC*Clay) + e*(OC*Sand)	Katterer et al., 2006 (49)	a: 1.319 b: 0.001 c: 0.071 d: 0.0002 e: 0.0005	a: 1.335 b: -0.003 c: 0.056 d: 0.0001 e: -0.0006

	BD = exp(a – b*OC + c*Clay – d*(Clay^2) – e*Silt)	Kaur et al., 2002 (50)	a: 0.334 b: 0.099 c: 0.005 d: 0.0001 e: 0.0005	a: 0.509 b: 0.102 c: 0.004 d: 0.00001 e: 0.003
	BD = a – b*Clay – c*Silt – d*Sand – e*OC	Makovniková et al., 2017 (53)	a: 7.080 b: 0.057 c: 0.057 d: 0.057 e: 0.084	a: 1.302 b: -0.006 c: 0.0014 d: -0.003 e: 0.076
	BD = a – b*ln(Sand) – c*ln(OC)	Sevastas et al., 2018 (67)	a: 1.191 b: 0.016 c: 0.187	a: 1.043 b: -0.078 c: 0.215
	BD = a – b*OC – c*Silt - d*Clay	Tomasella and Hodnett, 1998 (72)	a: 1.3470 b: 0.0842 c: -0.0004 d: 0.00005	a: 1.591 b: 0.076 c: 0.004 d: -0.003
	BD = a + b*Depth – c*OC	Yanti et al., 2021 (73)	a: 1.268 b: 0.0011 c: 0.1774	a: 1.307 b: 0.003 c: 0.060
H **With only texture, pH or depth**	BD = a – b*Clay – c*pH	Barros et al 2015 (11)	a: 0.7 b: 0.002 c: 0.074	a: 0.330 b: 0.005 c:-0.180
	BD = a – b*Clay + c*Silt	Heinonen 1977 (29)	a: 1.177 b: 0.001 c: 0.0004	a: 1.544 b: -6.878 c: -0.006
	BD = a – b*Clay – c*Sand – d*Silt	Makovníková et al., 2017 (52)	a: 2.636 b: 0.016 c: 0.015 d: 0.014	a: 1.073 b: -0.005 c: -0.005 d: 0.001

	BD = a + b*Clay – c*log(Silt)	Akpa 2016 (4)	a: 1.1900 b: 0.0010 c: 0.0005	a: 1.671 b: 0.0002 c: 0.102
	BD = a + b*Sand – c*log(Silt) + d*Silt	Akpa 2016 (2), (3)	a: 1.2299 b: 0.0009 c: 0.0678 d: 0.0037	a: 1.486 b: 0.0002 c: -0.024 d: -0.007
	BD = a + b*Depth - c*(1/Clay) – d*(1/Depth)	Qiao et al., 2019 (60)	a: 1.173 b: 0.004 c: -0.052 d: 1.256	a: 1.129 b: 0.004 c: -0.372 d: 0.450
X **Could not NLS**	BD = a – (b*exp(c*OC)) – (d*log(Depth))	Hollis et al 2012 (30)	Error: stepfactor below minfactor	
	BD = (a + b*sin(c - d*OC) + (e*Clay)*sin(sin(f + g*pH - h*Clay)) - j*sin(sin(k + m*pH - n*Clay)))	Pereira et al., 2016 (57)	Error: arguments imply differing number of rows	

Table 2.6. Results of PTFs tested on Ontario (All Variables) and recalibrated using NLS. For both the Literature coefficients and the NLS generated coefficients, accuracy metrics of R^2, CCC and RMSE were generated.

Model Group	Reference	Original			NLS		
		R^2	CCC	RMSE	R^2	CCC	RMSE
A (Linear)	Katterer et al. 2006 (42)	0.45	0.40	0.32	0.45	0.62	0.20
	Katterer et al. 2006 (43)	0.45	0.51	0.29	0.45	0.62	0.20
BD = a – b*OC	Kobal et al., 2011 (51) (OC < 3.6%)	0.45	0.67	0.23	0.45	0.62	0.20
	Kobal et al., 2011 (51) (OC > 3.6%)	0.45	0.34	0.26	0.45	0.62	0.20
	Manrique and Jones 1991 (54)	0.45	0.63	0.22	0.45	0.62	0.20
	Saini 1966 (OM) (64)	0.45	0.47	0.30	0.45	0.62	0.20
	Saini 1966 (OM) (65)	0.45	0.47	0.30	0.45	0.62	0.20
	Saini 1966 (OM) (66)	0.45	0.47	0.30	0.45	0.62	0.20
B (Radical root)	Alexander 1980 (5)	0.50	0.64	0.20	0.50	0.67	0.19
	Alexander 1980 (6)	0.51	0.54	0.24	0.51	0.67	0.19
BD = a – b*OC$^{0.5}$	De Vos 2005 (20) (OM)	0.51	0.30	0.35	0.51	0.67	0.19
	Manrique and Jones 1991 (55)	0.51	0.66	0.19	0.50	0.67	0.19
	Tamminen and Starr 1994 (OM) (71)	0.51	0.67	0.19	0.51	0.67	0.19
C (Reciprocal)	Honeysett and Ratkowsky 1989 (OM) (35)	0.51	0.58	0.26	0.51	0.67	0.19
	Drew, 1973 (21) (OM)	0.51	0.52	0.25	0.51	0.67	0.19
BD = 1 / (a + b*OC)	Nanko et al. 2014 (56)	0.51	0.33	0.34	0.51	0.67	0.19
D (Multiple ln terms)	Curtis and Post 1964 (19) (OM)	0.39	0.18	1.65	0.49	0.65	0.19
	Federer 1983 (22) (OM)	0.25	0.02	1.13	0.49	0.65	0.19
ln(BD) = a + b*ln(OC) + c*[ln(OC)]^2	Huntington et al. 1989 (39 OM)	0.21	0.02	1.15	0.49	0.65	0.19
	Huntington et al. 1989 (40 OC)	0.44	0.63	0.21	0.49	0.65	0.19
	Prévost 2004 (59)	0.29	0.03	1.09	0.49	0.65	0.19
E (Natural exponent)	Grigal et al. 1989 (24) (OM)	0.50	0.19	0.78	0.51	0.67	0.19
	Grigal et al. 1989 (25) (OM)	0.47	0.68	0.20	0.51	0.67	0.19

Model Group	Reference	Original			NLS		
		R^2	CCC	RMSE	R^2	CCC	RMSE
BD = a + b*exp^(c*OC)	Grigal et al., 1989 (26) (OM) (Peat 25-175 cm)	0.49	0.03	2.49	0.51	0.67	0.19
	Grigal et al. 1989 (27) (OM)	0.50	0.65	0.19	0.51	0.67	0.19
	Hossain et al. 2015 (36)	0.50	0.65	0.21	0.51	0.67	0.19
	Hossain et al. 2015 (37)	0.49	0.08	1.17	0.51	0.67	0.19
	Hossain et al. 2015 (38)	0.49	0.58	0.20	0.51	0.67	0.19
BD = a*exp(b*OC)	Adelbaki 2018 (1)	0.47	0.26	0.29	0.50	0.67	0.19
	Grigal et al., 1989 (26) (OM) (Peat 0-25 cm)	0.49	0.03	2.45	0.50	0.67	0.19
	Song et al., 2005 (69)	0.45	0.05	0.31	0.50	0.67	0.19
	Song et al., 2005 (70)	0.45	0.04	0.32	0.50	0.67	0.19
BD = a*exp(b*$OC^{0.5}$)	Alexander 1989 (7)	0.51	0.18	0.44	0.50	0.66	0.19
	Alexander 1989 (8)	0.51	0.08	0.76	0.50	0.66	0.19
	Alexander 1989 (9)	0.51	0.17	0.46	0.50	0.66	0.19
	Alexander 1989 (10)	0.51	0.23	0.37	0.50	0.66	0.19
BD = a*exp(b*OC)	Ruehlmann and Körschens 2009 (61)	0.46	0.04	0.44	0.50	0.67	0.19
	Ruehlmann and Körschens 2009 (62)	0.46	0.02	0.70	0.50	0.67	0.19
	Ruehlmann and Körschens 2009 (63)	0.46	0.12	0.26	0.50	0.67	0.19
BD = exp(a – b*$OM^{0.5}$)	Han et al 2012 (OM) (28)	0.51	0.14	0.42	0.50	0.66	0.19
F (With only OM/OC) BD = [a – b*$OC^{0.06}$]$^{3.85}$	Beutler et al., 2017 (15)	0.42	0.02	5.98	0.45	0.61	0.20
BD = a/OC*100	Gosselink et al., 1984 (23)	0.19	0.05	4.68	0.26	0.14	0.95

Model Group	Reference	Original			NLS		
		R^2	CCC	RMSE	R^2	CCC	RMSE
BD = a – b*ln(OC)	Hollis et al., 2012 (34)	0.46	0.56	0.30	0.46	0.63	0.19
BD = a – b*log(OM)	Jeffrey, 1970 (OM) (41)	0.46	0.38	0.62	0.46	0.63	0.19
BD = (a – b*OM^c)^d	Premrov et al., 2018 (58)	0.46	0.07	1.10	0.46	0.50	0.21
BD = a – b*OC + c*OC^2	Sevastas et al., 2018 (68)	0.09	-0.09	1.65	0.51	0.67	0.19
G (With OC/OM, other terms) BD = a – b*Clay – c*OC – d*pH + e*Sand	Bernoux et al., 1998 (12)	0.21	0.34	0.26	0.46	0.63	0.19
BD = a*Sand – b*OC	Bernoux et al., 1998 (13)	0.07	0.11	0.77	0.03	-0.11	0.63
BD = [a – b*(Clay+1)$^{0.46}$ – c*OC$^{0.55}$]$^{1.33}$	Beutler et al., 2017 (14)	0.29	0.05	0.59	0.51	0.67	0.19
BD = a – b*Clay – c*Sand – d*OC	Botula et al., 2015 (16)	0.25	0.08	0.38	0.45	0.62	0.20
BD = a – b*OC – c*Clay – d*Sand + e*pH	Brahim et al., 2012 (17)	0.44	0.54	0.28	0.46	0.63	0.19
BD = a – b*OC + c*Sand + d*Silt + e*pH	Brahim et al., 2012 (18)	0.35	0.23	0.48	0.47	0.63	0.19
BD = a + (b*exp^(c*OC)) + d*Sand – e*Clay	Hollis et al., 2012 (31)	0.47	0.52	0.26	0.51	0.67	0.19
BD = a + (b*exp^(c*OC)) + d*Sand – e*Clay	Hollis et al., 2012 (33)	0.49	0.62	0.19	0.51	0.67	0.19
BD = a – (b*ln(OC)) + (c*ln(Depth)) + (d*Sand)	Hollis et al., 2012 (32)	0.37	0.43	0.26	0.47	0.64	0.19
BD = a – b*OC + c*(OC*Clay)	Katterer et al., 2006 (44)	0.19	-0.05	3.93	0.46	0.63	0.20
BD = a + b*Clay – c*OC – d*(OC*Clay)	Katterer et al., 2006 (45)	0.26	0.04	6.56	0.47	0.63	0.19
BD = a + b*(Sand + CF) – c*OC	Katterer et al., 2006 (46)	0.01	0.003	6.49	0.46	0.62	0.20
BD = a + b*(Sand + CF) – c*OC	Katterer et al., 2006 (47)	0.01	0.01	5.31	0.46	0.62	0.20

Model Group	Reference	Original			NLS		
		R^2	CCC	RMSE	R^2	CCC	RMSE
BD = a – b*OC + c*(OC*Sand)	Katterer et al., 2006 (48)	0.27	-0.03	6.10	0.46	0.63	0.19
BD = a – b*Sand – c*OC – d*(OC*Clay) + e*(OC*Sand)	Katterer et al., 2006 (49)	0.15	-0.01	12.48	0.47	0.63	0.19
BD = exp(a – b*OC + c*Clay – d*(Clay^2) – e*Silt)	Kaur et al., 2002 (50)	0.43	0.46	0.67	0.51	0.68	0.19
BD = a – b*Clay – c*Silt – d*Sand – e*OC	Makovnikova et al., 2017 (53)	0.41	0.31	0.37	0.45	0.62	0.20
BD = a – b*ln(Sand) – c*ln(OC)	Sevastas et al., 2018 (67)	0.35	0.37	0.45	0.46	0.63	0.19
BD = a – b*OC – c*Silt - d*Clay	Tomasella and Hodnett, 1998 (72)	0.21	0.42	0.24	0.45	0.62	0.20
BD = a + b*Depth – c*OC	Yanti et al., 2021 (73)	0.45	0.46	0.40	0.46	0.62	0.20
H (only texture, pH or depth) BD = a – b*Clay – c*pH	Barros et al 2015 (11)	0.001	-0.01	0.56	0.07	0.13	0.26
BD = a – b*Clay + c*Silt	Heinonen 1977 (29)	0.002	0.01	0.39	0.01	0.01	0.26
BD = a – b*Clay – c*Sand – d*Silt	Makovníková et al., 2017 (52)	0.003	-0.01	0.55	0.008	0.01	0.27
BD = a + b*Depth - c*(1/Clay) – d*(1/Depth)	Qiao et al., 2019 (60)	0.46	0.07	1.1	0.18	0.30	0.24
BD = 1.177 + 0.00263*Sand - 0.0439*log(Silt) + 0.00208*Silt	Akpa 2016 (2)	0.003	0.02	0.27	0.01	0.01	0.26
	Akpa 2016 (3)	0.002	0.02	0.27	0.01	0.01	0.26
BD = 1.512–0.00322*Clay-0.0865*log(Silt)	Akpa 2016 (4)	0.003	0.03	0.28	0.01	0.01	0.26

RMSE root mean square error; R^2 prediction coefficient of determination; CCC concordance correlation coefficient.

Table 2.7. Results of PTFs tested on BC (All Variables) and recalibrated using NLS. For both the Literature coefficients and the NLS generated coefficients, accuracy metrics of R^2, CCC and RMSE were generated.

Model Group	Reference	Original			NLS		
		R^2	CCC	RMSE	R^2	CCC	RMSE
A (Linear) **BD = a – b*OC**	Katterer et al. 2006 (42)	0.31	0.47	0.34	0.43	0.40	0.33
	Katterer et al. 2006 (43)	0.31	0.53	0.34	0.40	0.39	0.32
	Kobal et al., 2011 (51) (OC <3.6%)	0.31	0.53	0.38	0.40	0.39	0.33
	Kobal et al., 2011 (51) (OC > 3.6%)	0.31	0.20	0.44	0.40	0.39	0.33
	Manrique and Jones 1991 (54)	0.31	0.53	0.33	0.41	0.39	0.33
	Saini 1966 (OM) (64)	0.31	0.50	0.34	0.41	0.39	0.33
	Saini 1966 (OM) (65)	0.31	0.50	0.34	0.41	0.39	0.33
	Saini 1966 (OM) (66)	0.31	0.50	0.34	0.41	0.39	0.33
B (Radical root) **BD = a – b*OC$^{0.5}$**	Alexander 1980 (5) (OC)	0.42	0.56	0.29	0.47	0.53	0.30
	Alexander 1980 (6) (OC)	0.42	0.54	0.30	0.45	0.52	0.29
	De Vos 2005 (20) (OM)	0.42	0.40	0.35	0.46	0.52	0.29
	Manrique and Jones 1991 (55)	0.42	0.57	0.30	0.46	0.53	0.29
	Tamminen and Starr 1994 (OM) (71)	0.42	0.52	0.32	0.33	0.14	0.41
C (Reciprocal) **BD = 1 / (a + b*OC)**	Honeysett and Ratkowsky 1989 (OM) (35)	0.44	0.61	0.31	0.47	0.57	0.29
	Drew, 1973 (21) (OM)	0.41	0.53	0.30	0.49	0.58	0.28
	Nanko et al. 2014 (56)	0.43	0.26	0.50	0.45	0.55	0.29
D (Multiple ln terms) **lnBD = a + b*ln(OC) + c*[ln(OC)]^2**	Curtis and Post 1964 (19)	0.32	0.21	1.79	0.48	0.57	0.28
	Federer 1983 (22)	0.29	0.02	1.28	0.48	0.57	0.28
	Huntington et al. 1989 (39) (OM)	0.25	0.02	1.30	0.48	0.56	0.28
	Huntington et al. 1989 (40) (OC)	0.37	0.48	0.34	0.48	0.58	0.28
	Prévost 2004 (59) (OM)	0.33	0.04	1.23	0.47	0.57	0.28
E	Grigal et al. 1989 (24) (OM)	0.41	0.27	0.76	0.47	0.57	0.28

Model Group	Reference	Original			NLS		
		R^2	CCC	RMSE	R^2	CCC	RMSE
(Natural exponent)	Grigal et al. 1989 (25) (OM)	0.46	0.61	0.30	0.47	0.57	0.28
	Grigal et al. 1989 (26) (OM) (Peat 25-175cm)	0.45	0.04	2.47	0.47	0.57	0.28
BD = a + b*exp^(c*OC)	Grigal et al. 1989 (27) (OM)	0.40	0.48	0.34	0.47	0.56	0.29
	Hossain et al. 2015 (36)	0.45	0.61	0.28	0.47	0.57	0.28
	Hossain et al. 2015 (37)	0.38	0.12	1.10	0.47	0.57	0.29
	Hossain et al. 2015 (38)	0.38	0.44	0.32	0.48	0.57	0.28
BD = a*exp(b*OC)	Adelbaki 2018 (1)	0.34	0.27	0.34	0.45	0.52	0.30
	Grigal et al., 1989 (26) (OM) (Peat 0-25 cm)	0.46	0.61	0.30	0.45	0.53	0.30
	Song et al., 2005 (69)	0.31	0.05	0.37	0.43	0.50	0.30
	Song et al., 2005 (70)	0.31	0.05	0.37	0.44	0.52	0.30
BD = a*exp(b*$OC^{0.5}$)	Alexander 1989 (7)	0.43	0.28	0.41	0.46	0.55	0.28
	Alexander 1989 (8)	0.43	0.14	0.69	0.48	0.57	0.28
	Alexander 1989 (9)	0.43	0.27	0.43	0.47	0.56	0.29
	Alexander 1989 (10)	0.43	0.32	0.36	0.47	0.57	0.28
BD = a*exp(-c*OC)	Ruehlmann and Korschens 2009 (61)	0.32	0.07	0.42	0.44	0.52	0.29
	Ruehlmann and Korschens 2009 (62)	0.31	0.03	0.61	0.44	0.52	0.29
	Ruehlmann and Korschens 2009 (63)	0.32	0.09	0.37	0.44	0.52	0.29
BD = exp(a – b*$OM^{0.5}$)	Han et al 2012 (28) (OM)	0.43	0.23	0.40	0.47	0.57	0.28
F **(With only OM/OC)** **BD = [a – b*$OC^{0.06}$]$^{3.85}$**	Beutler et al., 2017 (15)	0.46	0.03	5.66	0.47	0.56	0.29
BD = a/OC*100	Gosselink et al., 1984 (23)	0.23	0.05	5.73	0.32	0.23	0.94
BD = a – b*ln(OC)	Hollis et al., 2012 (34)	0.44	0.61	0.35	0.47	0.57	0.29
BD = a – b*log(OM)	Jeffrey, 1970 (41)	0.44	0.50	0.63	0.48	0.56	0.29
BD = (a – b*OM^{c})^{d}	Premrov et al., 2018 (58)	0.44	0.10	1.16	0.47	0.44	0.31

Model Group	Reference	Original			NLS		
		R^2	CCC	RMSE	R^2	CCC	RMSE
BD = a – b*OC + c*OC²	Sevastas et al., 2018 (68)	0.03	-0.05	2.44	0.45	0.54	0.29
G (With OC/OM, other terms) BD = a – b*Clay – c*OC – d*pH + e*Sand	Bernoux et al., 1998 (12)	0.24	0.28	0.38	0.42	0.48	0.32
BD = a*Sand – b*OC	Bernoux et al., 1998 (13)	0.21	0.24	0.80	0.11	0.15	0.65
BD = [a – b*(Clay+1)$^{0.46}$ – c*OC$^{0.55}$]$^{1.33}$	Beutler et al., 2017 (14)	0.41	0.11	0.50	0.45	0.53	0.29
BD = a – b*Clay – c*Sand – d*OC	Botula et al., 2015 (16)	0.09	0.06	0.40	0.42	0.44	0.32
BD = a – b*OC – c*Clay – d*Sand + e*pH	Brahim et al., 2012 (17)	0.25	0.48	0.36	0.42	0.48	0.31
BD = a – b*OC + c*Sand + d*Silt + e*pH	Brahim et al., 2012 (18)	0.31	0.34	0.45	0.41	0.47	0.32
BD = a + (b*exp^(c*OC)) + d*Sand – e*Clay	Hollis et al., 2012 (31)	0.46	0.57	0.30	0.48	0.58	0.28
BD = a + (b*exp^(c*OC)) + d*Sand – e*Clay	Hollis et al., 2012 (33)	0.46	0.53	0.30	0.48	0.58	0.28
BD = a – (b*ln(OC)) + (c*ln(Depth)) + (d*Sand)	Hollis et al., 2012 (32)	0.45	0.54	0.30	0.48	0.57	0.28
BD = a – b*OC + c*(OC*Clay)	Katterer et al., 2006 (44)	0.27	-0.13	2.53	0.41	0.39	0.34
BD = a + b*Clay – c*OC – d*(OC*Clay)	Katterer et al., 2006 (45)	0.29	0.08	4.64	0.39	0.37	0.34
BD = a + b*(Sand + CF) – c*OC	Katterer et al., 2006 (46)	0.22	0.02	9.57	0.45	0.51	0.30
BD = a + b*(Sand + CF) – c*OC	Katterer et al., 2006 (47)	0.24	0.02	7.85	0.44	0.50	0.30
BD = a – b*OC + c*(OC*Sand)	Katterer et al., 2006 (48)	0.16	-0.05	5.13	0.41	0.39	0.33
BD = a – b*Sand – c*OC – d*(OC*Clay) + e*(OC*Sand)	Katterer et al., 2006 (49)	0.14	-0.03	9.78	0.42	0.43	0.33

Model Group	Reference	Original			NLS		
		R^2	CCC	RMSE	R^2	CCC	RMSE
BD = exp(a – b*OC + c*Clay – d*(Clay^2) – e*Silt)	Kaur et al., 2002 (50)	0.44	0.47	0.42	0.42	0.50	0.32
BD = a – b*Clay – c*Silt – d*Sand – e*OC	Makovnikova et al., 2017 (53)	0.27	0.40	0.36	0.13	0.14	0.37
BD = a – b*ln(Sand) – c*ln(OC)	Sevastas et al., 2018 (67)	0.33	0.45	0.45	0.49	0.58	0.28
BD = a – b*OC – c*Silt - d*Clay	Tomasella and Hodnett, 1998 (72)	0.29	0.44	0.35	0.42	0.44	0.32
BD = a + b*Depth – c*OC	Yanti et al., 2021 (73)	0.33	0.42	0.54	0.43	0.49	0.31
H (only texture, pH or depth) BD = a – b*Clay – c*pH	Barros et al 2015 (11)	0.01	-0.02	0.61	0.20	0.23	0.36
BD = a – b*Clay + c*Silt	Heinonen 1977 (29)	0.10	-0.06	0.42	0.16	0.16	0.36
BD = a – b*Clay – c*Sand – d*Silt	Makovníková et al., 2017 (52)	0.05	-0.07	0.48	0.15	0.16	0.36
BD = a + b*Depth - c*(1/Clay) – d*(1/Depth)	Qiao et al., 2019 (60)	0.03	-0.18	0.62	0.32	0.40	0.33
BD = a + b*Sand – c*log(Silt) + d*Silt	Akpa 2016 (2)	0.07	0.09	0.38	0.13	0.13	0.37
	Akpa 2016 (3)	0.09	0.11	0.38	0.14	0.14	0.37
BD = a – b*Clay – c*log(Silt)	Akpa 2016 (4)	0.07	0.12	0.40	0.17	0.11	0.38

RMSE root mean square error; R^2 prediction coefficient of determination; CCC concordance correlation coefficient.

Table 2.8. Results of PTFs tested on BC (C and BD) and recalibrated using NLS. For both the Literature coefficients and the NLS generated coefficients, accuracy metrics of R^2, CCC and RMSE were generated.

Model Group	Reference	Original			NLS		
		R^2	CCC	RMSE	R^2	CCC	RMSE
A (Linear) BD = a – b*OC	Katterer et al. 2006 (42)	0.23	0.43	0.36	0.24	0.36	0.34
	Katterer et al. 2006 (43)	0.23	0.47	0.38	0.24	0.36	0.34
	Kobal et al., 2011 (51) (OC < 3.6%)	0.23	0.44	0.43	0.24	0.36	0.34
	Kobal et al., 2011 (51) (OC > 3.6%)	0.25	0.18	0.46	0.24	0.36	0.34
	Manrique and Jones 1991 (54)	0.23	0.46	0.37	0.24	0.36	0.34
	Saini 1966 (OM) (64)	0.23	0.45	0.36	0.24	0.36	0.34
	Saini 1966 (OM) (65)	0.23	0.43	0.35	0.24	0.35	0.34
	Saini 1966 (OM) (66)	0.23	0.46	0.36	0.24	0.35	0.34
B (Radical root) BD = a – b*OC^0.5	Alexander 1980 (5)	0.33	0.50	0.32	0.33	0.48	0.32
	Alexander 1980 (6)	0.33	0.50	0.32	0.33	0.48	0.32
	De Vos 2005 (20) (OM)	0.33	0.38	0.36	0.33	0.48	0.32
	Manrique and Jones 1991 (55)	0.33	0.50	0.33	0.33	0.48	0.32
	Tamminen and Starr 1994 (OM) (71)	0.33	0.45	0.35	0.33	0.48	0.32
C (Reciprocal) BD = 1 / (a + b*OC)	Honeysett and Ratkowsky 1989 (OM) (35)	0.33	0.55	0.33	0.32	0.49	0.32
	Drew, 1973 (21) (OM)	0.31	0.47	0.32	0.32	0.49	0.32
	Nanko et al. 2014 (56)	0.32	0.23	0.53	0.32	0.49	0.32
D (Multiple ln terms) lnBD = a + b*ln(OC) + c*[ln(OC)]^2	Curtis and Post 1964 (19) (OM)	0.15	0.14	1.98	0.39	0.54	0.30
	Federer 1983 (22) (OM)	0.27	0.03	1.29	0.39	0.54	0.30
	Huntington et al. 1989 (39 OM)	0.24	0.03	1.30	0.39	0.55	0.30
	Huntington et al. 1989 (40 OC)	0.20	0.35	0.40	0.39	0.54	0.30
	Prévost 2004 (59)	0.31	0.04	1.24	0.39	0.55	0.30

E (Natural exponent) Db = a + b*exp^(c*OC)	Grigal et al. 1989 (24) (OM)	0.30	0.26	0.75	0.39	0.54	0.31
	Grigal et al. 1989 (25) (OM)	0.36	0.53	0.34	0.39	0.54	0.30
	Grigal et al., 1989 (26) (OM) (Peat 0-25 cm)	0.28	0.04	2.38	0.28	0.44	0.33
	Grigal et al. 1989 (27) (OM)	0.29	0.40	0.38	0.39	0.54	0.31
	Hossain et al. 2015 (36)	0.35	0.54	0.31	0.39	0.54	0.31
	Hossain et al. 2015 (37)	0.27	0.12	1.08	0.39	0.54	0.30
	Hossain et al. 2015 (38)	0.27	0.38	0.35	0.39	0.54	0.30
Db = a*exp(b*OC)	Abdelbaki 2018 (1)	0.25	0.25	0.35	0.28	0.44	0.33
	Grigal et al., 1989 (26) (OM) (Peat 0-25 cm)	0.29	0.04	2.42	0.39	0.54	0.31
	Song et al., 2005 (69)	0.23	0.04	0.38	0.28	0.44	0.33
	Song et al., 2005 (70)	0.23	0.04	0.38	0.28	0.44	0.33
Db = a*exp(b*OC^0.5)	Alexander 1989 (7)	0.34	0.27	0.41	0.36	0.52	0.31
	Alexander 1989 (8)	0.34	0.14	0.68	0.37	0.53	0.31
	Alexander 1989 (9)	0.34	0.26	0.42	0.37	0.53	0.31
	Alexander 1989 (10)	0.34	0.31	0.36	0.37	0.53	0.31
Db = (a – b*c)*exp(-c*OC)	Ruehlmann and Korschens 2009 (61)	0.23	0.07	0.42	0.28	0.44	0.33
	Ruehlmann and Korschens 2009 (62)	0.23	0.03	0.61	0.28	0.44	0.33
	Ruehlmann and Korschens 2009 (63)	0.23	0.08	0.38	0.28	0.44	0.33
BD = e^(a – b*OM^0.5)	Han et al 2012 (28)	0.34	0.22	0.40	0.37	0.53	0.31
F (With only OM/OC) BD = $[a – b*OC^{0.06}]^{3.85}$	Beutler et al., 2017 (15)	0.35	0.03	8.31	0.37	0.40	0.35
BD = a/OC*100	Gosselink et al., 1984 (23)	0.31	0.47	0.32	0.22	0.13	1.17
BD = a – b*ln(OC)	Hollis et al., 2012 (34)	0.38	0.59	0.36	0.39	0.54	0.31
BD = a – b*ln(OM)	Jeffrey, 1970 (41) (OM)	0.38	0.43	0.71	0.39	0.54	0.31
BD = $(a – b*OM^{c})^{d}$	Premrov et al., 2018 (58)	0.20	0.10	1.25	0.37	0.20	0.39
BD = $a – b*OC + c*OC^2$	Sevastas et al., 2018 (68)	0.04	-0.06	2.55	0.32	0.47	0.32

RMSE root mean square error; R^2 prediction coefficient of determination; CCC concordance correlation coefficient.

Table 2.9. Comparison of RMSE values for PTFs included in this study, with the RMSE values reported in the original paper ("Orig. study" column), and RMSE values produced when those PTFs were tested on regional datasets in selected studies.

PTF#	Original study	Ontario (All Variables)	BC (All Variables)	BC (C BD)	Abdelbaki 2018	Boschi 2018	Al-Qinna & Jaber 2013	De Vos 2005	Han 2012	Kaur 2002	Nanko 2014	Sevastas 2018	Botula et al., 2015	Vasiliniuc and Patriche 2015	Yi 2016
1	0.13	0.29	0.34	0.35								0.1406			
2	0.179	0.27	0.38												
3	0.152	0.27	0.38												
4	0.189	0.28	0.40												
5		0.20	0.29	0.32	0.16	0.19	0.326	0.30	0.14	0.19	0.319	0.1839		0.157	0.154
6		0.24	0.30	0.32	0.17				0.17	0.23	0.438	0.1390		0.171	0.143
7		0.44	0.41	0.41											
8		0.76	0.69	0.68											
9		0.46	0.43	0.42											
10		0.37	0.36	0.36											
11		0.56	0.61												
12		0.26	0.38										0.307		
13		0.77	0.80												
14	0.22	0.59	0.50												
15	0.26	5.98	5.66	8.31											
16	0.137	0.38	0.40									0.1673			
17		0.28	0.36												
18		0.48	0.45												
19		1.65	1.79	1.98	0.35	0.30				0.25	0.164	0.3362		0.263	0.276
20	0.16	0.35	0.35	0.36							0.540	0.1493			
21		0.25	0.30	0.32								0.1413			

PTF#	Original study	Ontario (All Variables)	BC (All Variables)	BC (C BD)	Abdelbaki 2018	Boschi 2018	Al-Qinna & Jaber 2013	De Vos 2005	Han 2012	Kaur 2002	Nanko 2014	Sevastas 2018	Botula et al., 2015	Vasiliniuc and Patriche 2015	Yi 2016
22		1.13	1.28	1.29	0.33	0.31	1.594	0.45	0.23	0.23	0.153	0.3865		0.261	0.295
23		4.68	5.73	0.32											
24		0.78	0.76	0.75											
25		0.20	0.30	0.34											
26		2.49	2.47	2.38											
27		0.19	0.34	0.38		0.23					0.187	0.2898			
28	0.13	0.42	0.40	0.40	0.23	0.24					0.276	0.1995		0.164	
29		0.39	0.42												
30	0.17	NA	NA												
31	0.13	0.26	0.30		0.57								0.207	0.198	
32	0.14	0.26	0.30												
33	0.15	0.19	0.30		0.16									0.163	
34	0.10	0.30	0.35	0.36	0.45							0.1535			
35		0.26	0.31	0.33				0.25			0.271	0.1517			
36		0.21	0.28	0.31								0.1741			
37		1.17	1.10	1.08											
38		0.20	0.32	0.35											
39		1.15	1.30	1.30	0.58	0.31	1.599	0.45	0.40	0.45		0.2737	0.503		0.337
40		0.21	0.34	0.40	0.27	0.24			0.18	0.20	0.209			0.213	0.230
41		0.62	0.63	0.71		0.25		0.34			0.204	0.3016			
42	0.13	0.32	0.34	0.36											
43	0.16	0.29	0.34	0.38											
44	0.14	3.93	2.53												
45	0.16	6.56	4.64												

PTF#	Original study	Ontario (All Variables)	BC (All Variables)	BC (C BD)	Abdelbaki 2018	Boschi 2018	Al-Qinna & Jaber 2013	De Vos 2005	Han 2012	Kaur 2002	Nanko 2014	Sevastas 2018	Botula et al., 2015	Vasiliniuc and Patriche 2015	Yi 2016
46	0.13	6.49	9.57												
47	0.16	5.31	7.85												
48	0.12	6.10	5.13												
49	0.16	12.48	9.78												
50	0.15	0.67	0.42		0.32	0.38	0.515	0.56	0.26			0.3972	0.362	0.371	0.545
51		0.23	0.38	0.43								0.2061			
52		0.55	0.48												
53		0.37	0.36												
54		0.22	0.33	0.37	0.24	0.17			0.35	0.20		0.1502		0.175	0.189
55		0.19	0.30	0.33	0.16	0.19	0.335	0.32	0.14	0.18	0.301		0.637	0.160	0.159
56	0.138	0.34	0.50	0.53	0.42										
57	0.123	NA	NA												
58	0.126	1.10	1.16	1.25											
59		1.09	1.23	1.24	0.23	0.23			0.15		0.190	0.2691		0.202	
60	0.079	1.1	0.62												
61	0.215	0.44	0.42	0.42		0.47						0.1335			
62	0.215	0.70	0.61	0.61	0.15	0.47	0.142								
63	0.215	0.26	0.37	0.38		0.47									
64		0.30	0.34	0.36											
65		0.30	0.34	0.35								0.1338			
66		0.30	0.34	0.36											
67	0.1195	0.45	0.45												
68	0.12449	1.65	2.44	2.55											
69		0.31	0.37	0.38								0.1844			

PTF#	Original study	Ontario (All Variables)	BC (All Variables)	BC (C BD)	Abdelbaki 2018	Boschi 2018	Al-Qinna & Jaber 2013	De Vos 2005	Han 2012	Kaur 2002	Nanko 2014	Sevastas 2018	Botula et al., 2015	Vasiliniuc and Patriche 2015	Yi 2016
70		0.32	0.37	0.38											
71		0.19	0.32	0.35	0.19	0.26		0.37			0.211	0.2813			
72		0.24	0.35		0.26	0.19	0.421		0.21	0.18		0.2156		0.205	0.254
73		0.40	0.54												

2.11. Supplementary Figures

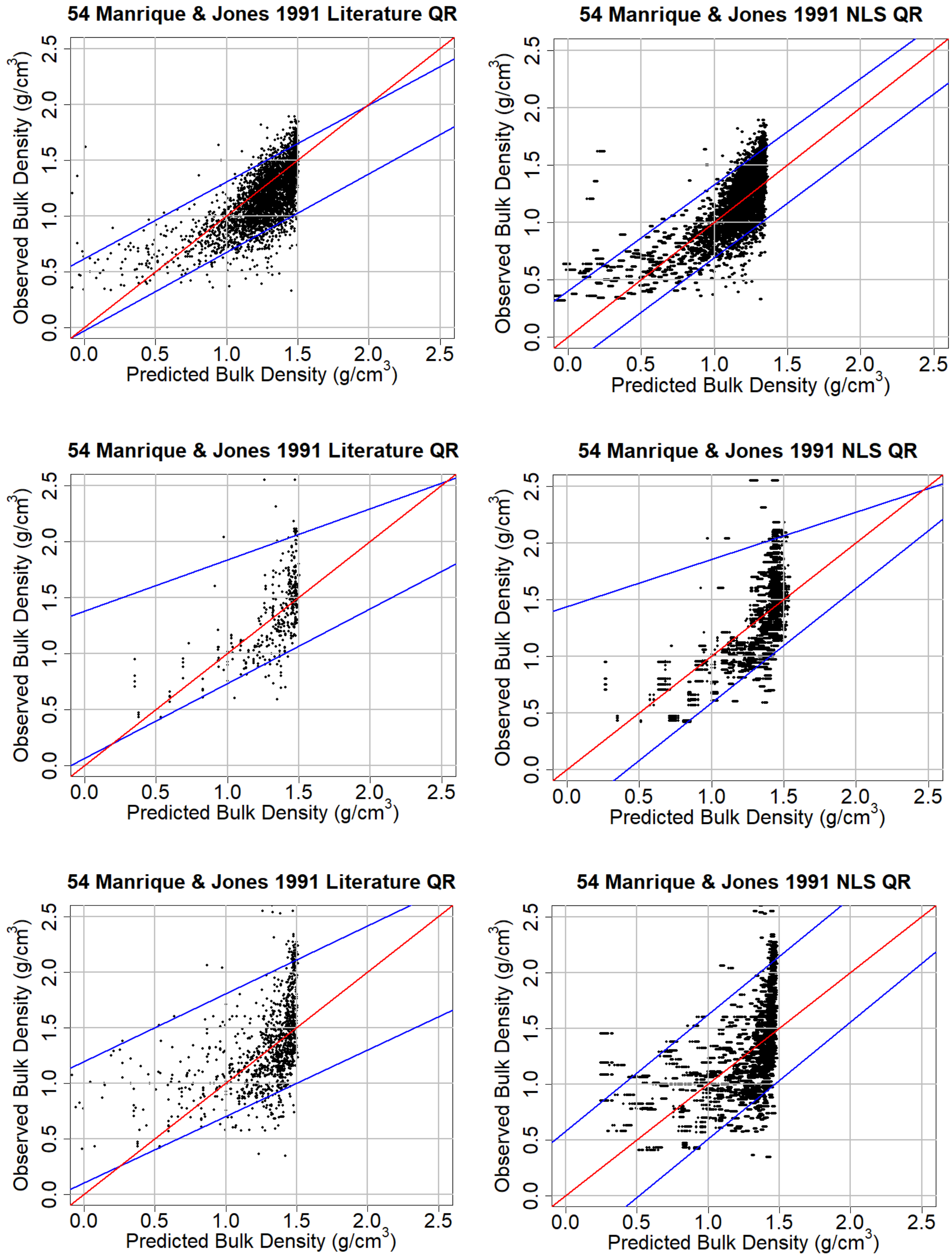
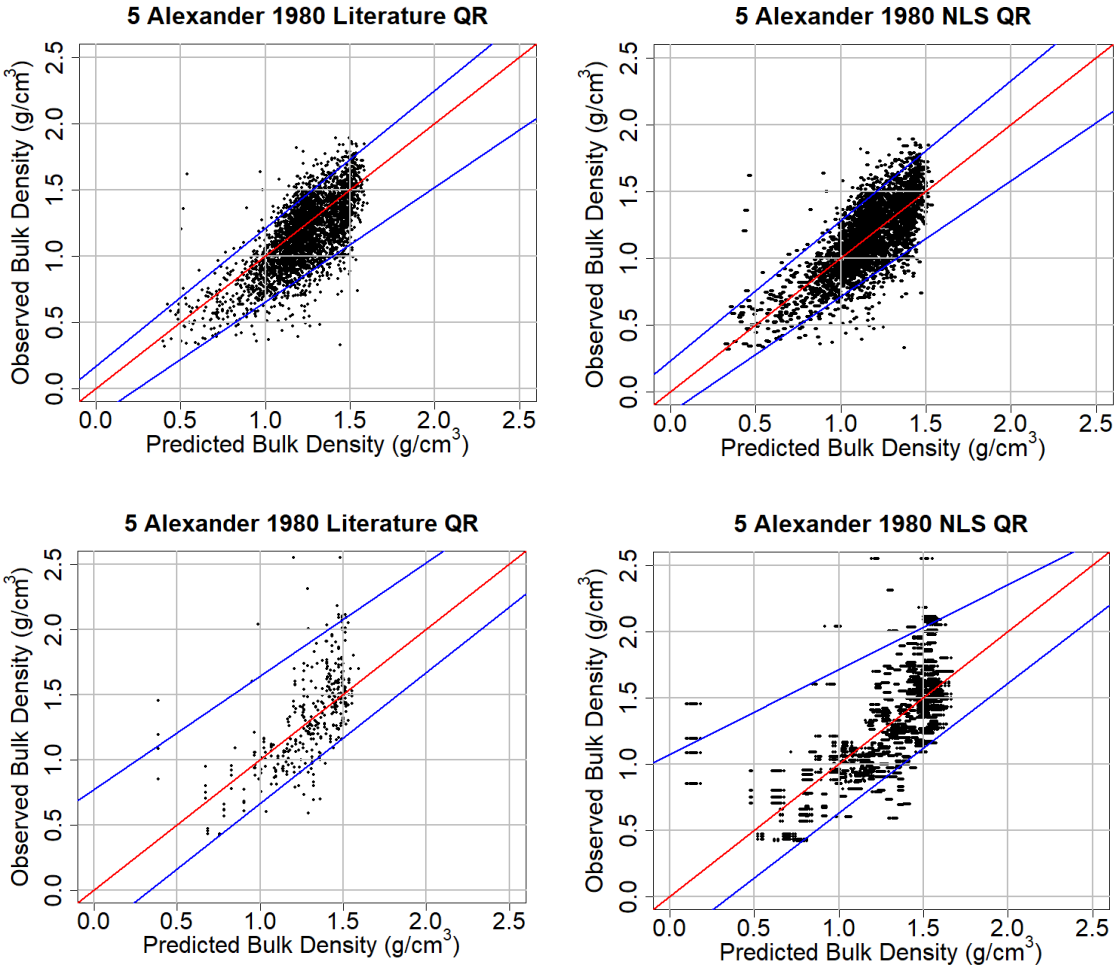


Figure 2.12. **Plots of observed versus predicted BD values for Group A (linear) functions. For each dataset used in this study, an example PTF was chosen for the model. This function was plotted using its literature coefficients, and then using its NLS-generated coefficients. Blue lines represent the 5th and 95th percentiles; red line is the 1:1 line. Ontario (All Variables) results are shown top left and right; BC (All Variables) are shown middle left and right; and BC (Carbon and Bulk Density) are shown bottom left and right.**



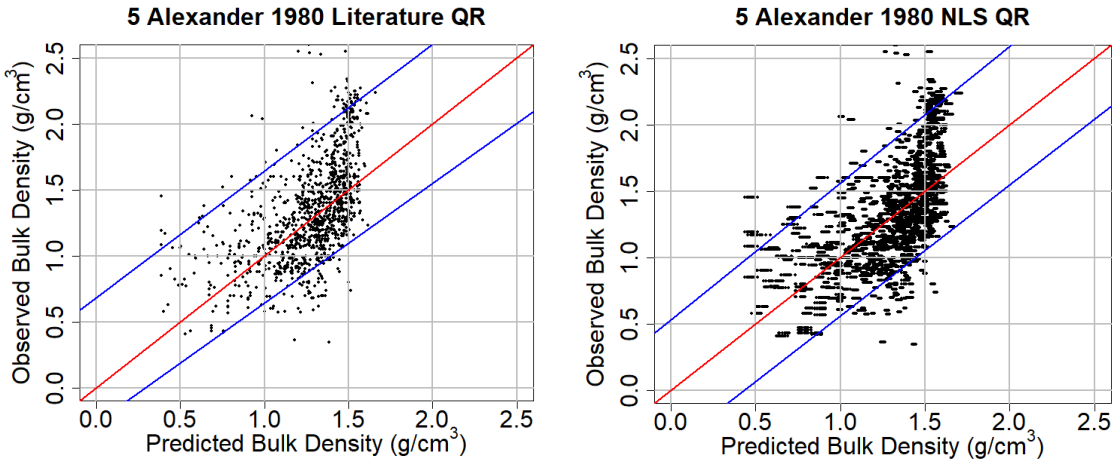
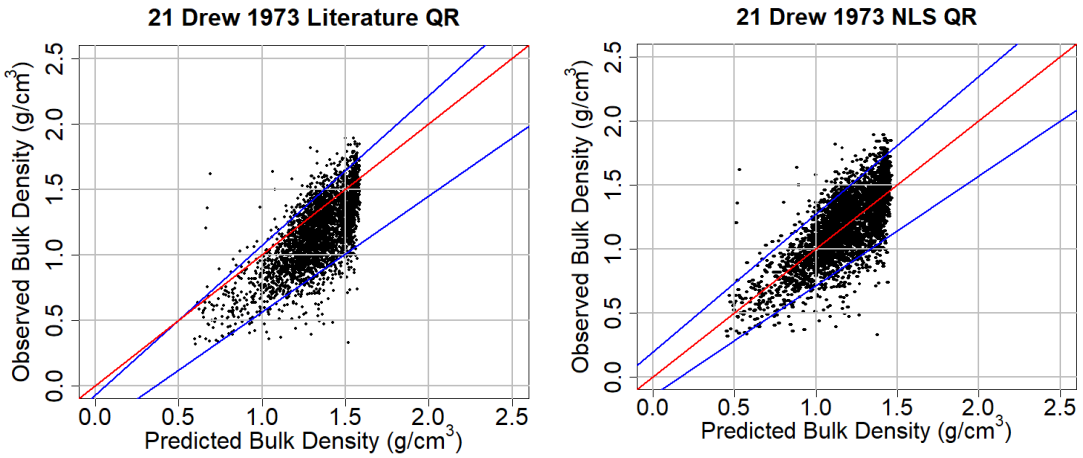


Figure 2.13. **Plots of observed versus predicted BD values for Group B (radical root) functions. For each dataset used in this study, an example PTF was chosen for the model. This function was plotted using its literature coefficients, and then using its NLS-generated coefficients. Blue lines represent the 5th and 95th percentiles; red line is the 1:1 line. Ontario (All Variables) results are shown top left and right; BC (All Variables) are shown middle left and right; BC (Carbon and Bulk Density) are shown bottom left and right.**



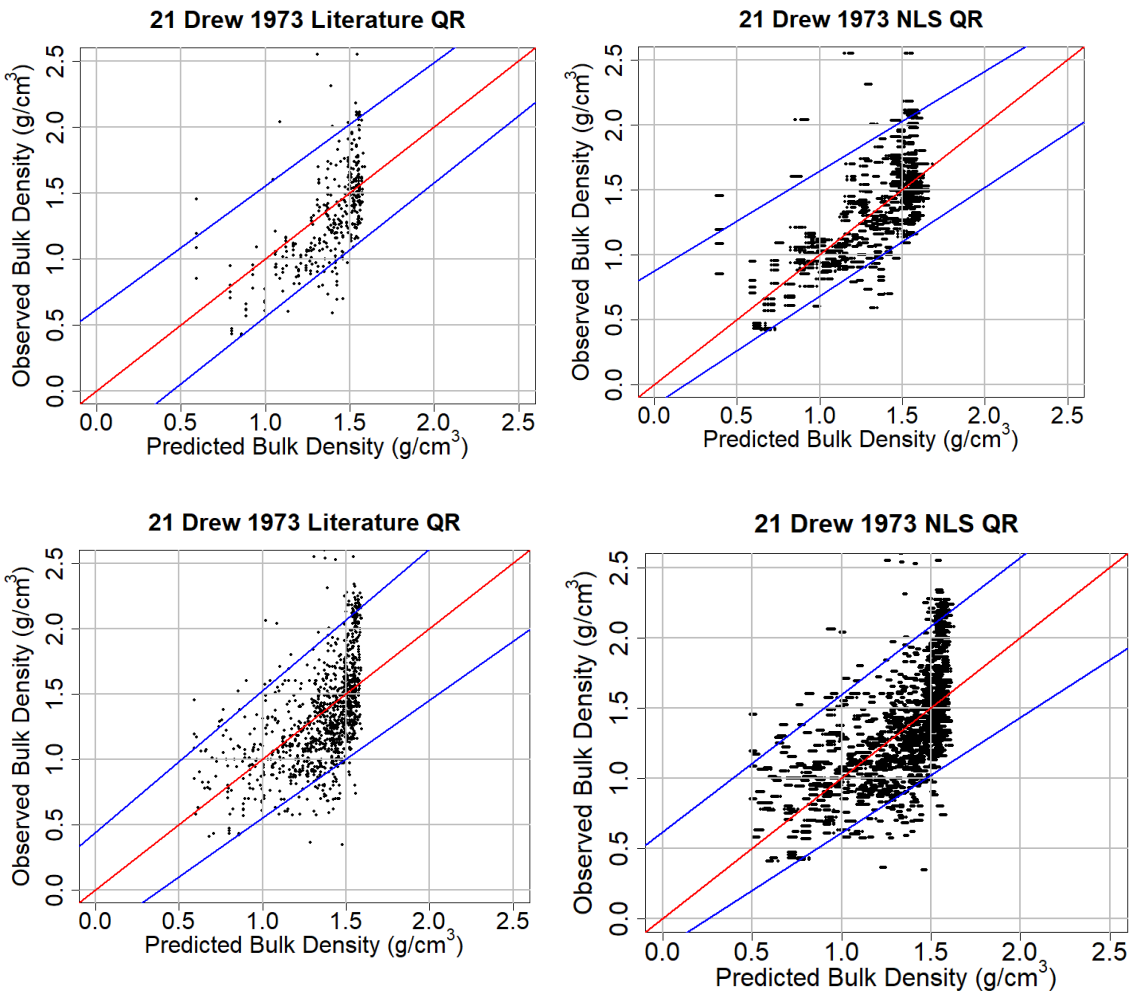
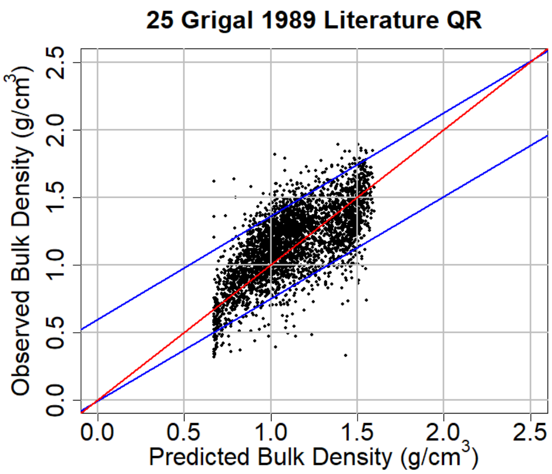
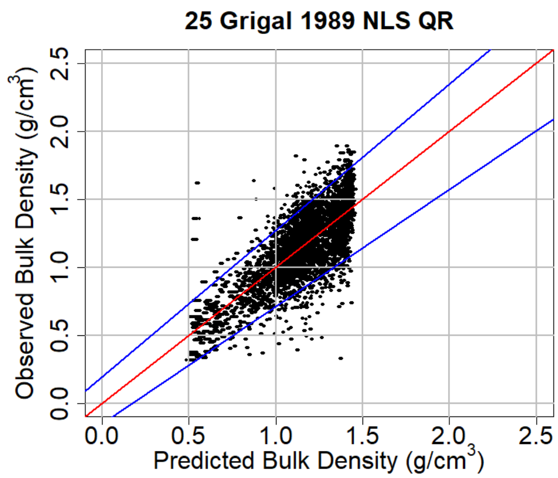
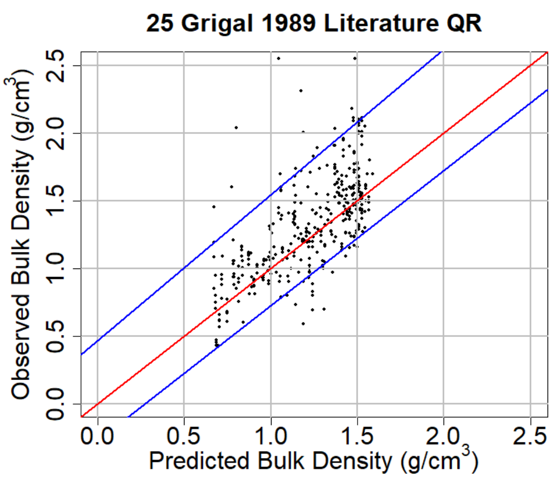
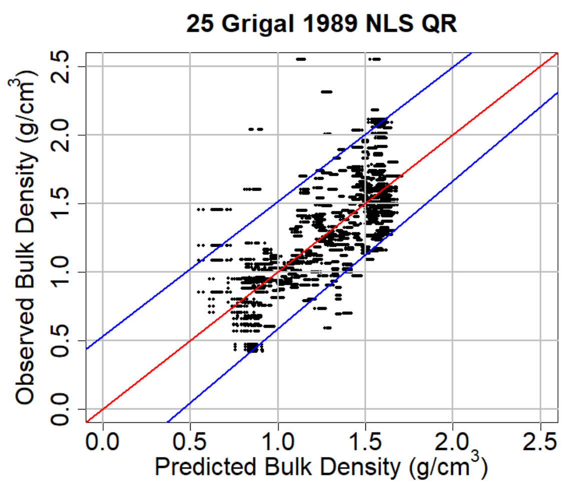


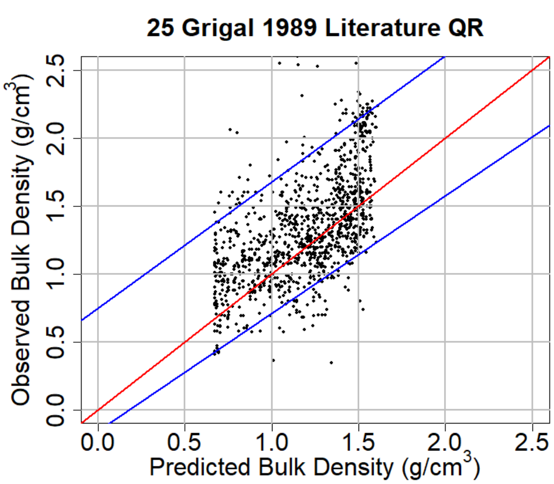
Figure 2.14. **Plots of observed versus predicted BD values for Group C (reciprocal) functions. For each dataset used in this study, an example PTF was chosen for the model. This function was plotted using its literature coefficients, and then using its NLS-generated coefficients. Blue lines represent the 5th and 95th percentiles; red line is the 1:1 line. Results for Ontario (All Variables) are shown top left and right; BC (All Variables) are shown middle left and right; and BC (Carbon and Bulk Density) are shown bottom left and right.**


25 Grigal 1989 Literature QR
Observed Bulk Density (g/cm³)
Predicted Bulk Density (g/cm³)
0.0
0.5
1.0
1.5
2.0
2.5


25 Grigal 1989 NLS QR
Observed Bulk Density (g/cm³)
Predicted Bulk Density (g/cm³)
0.0
0.5
1.0
1.5
2.0
2.5


25 Grigal 1989 Literature QR
Observed Bulk Density (g/cm³)
Predicted Bulk Density (g/cm³)
0.0
0.5
1.0
1.5
2.0
2.5


25 Grigal 1989 NLS QR
Observed Bulk Density (g/cm³)
Predicted Bulk Density (g/cm³)
0.0
0.5
1.0
1.5
2.0
2.5


25 Grigal 1989 Literature QR
Observed Bulk Density (g/cm³)
Predicted Bulk Density (g/cm³)
0.0
0.5
1.0
1.5
2.0
2.5

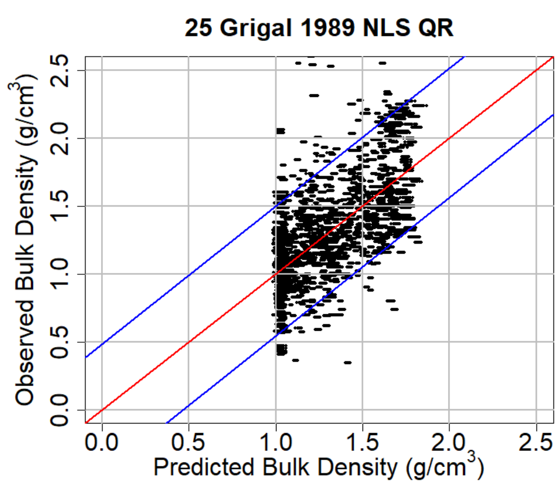

25 Grigal 1989 NLS QR
Observed Bulk Density (g/cm³)
Predicted Bulk Density (g/cm³)
0.0
0.5
1.0
1.5
2.0
2.5

Figure 2.15. **Plots of observed versus predicted BD values for Group E (natural exponent terms) functions. For each dataset used in this study, an example PTF was chosen for the model. This function was plotted using its literature coefficients, and then using its NLS-generated coefficients. Blue lines represent the 5th and 95th percentiles; red line is the 1:1 line. Results for Ontario (All Variables) are shown top left and right; BC (All Variables) are shown middle left and right; and BC (Carbon and Bulk Density) are shown bottom left and right.**

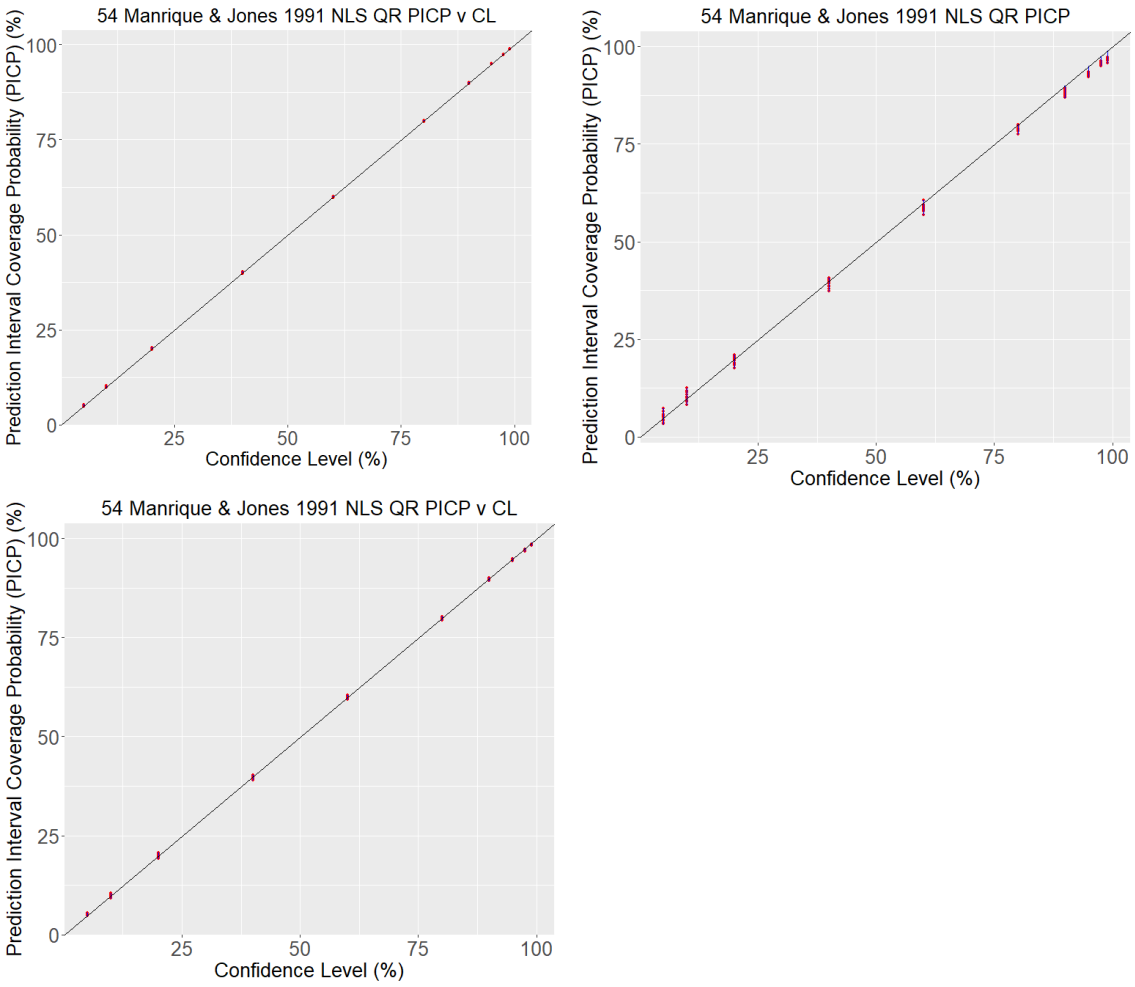


Figure 2.16. **PICP vs CL Graphs for each dataset. For Model Group A (linear functions), a representative PTF was chosen, and the PICP vs CL graph for that PTF is shown for each dataset. Ontario (All Variables) is shown top left; BC (All Variables) top right; and BC (Carbon and Bulk Density) bottom left.**

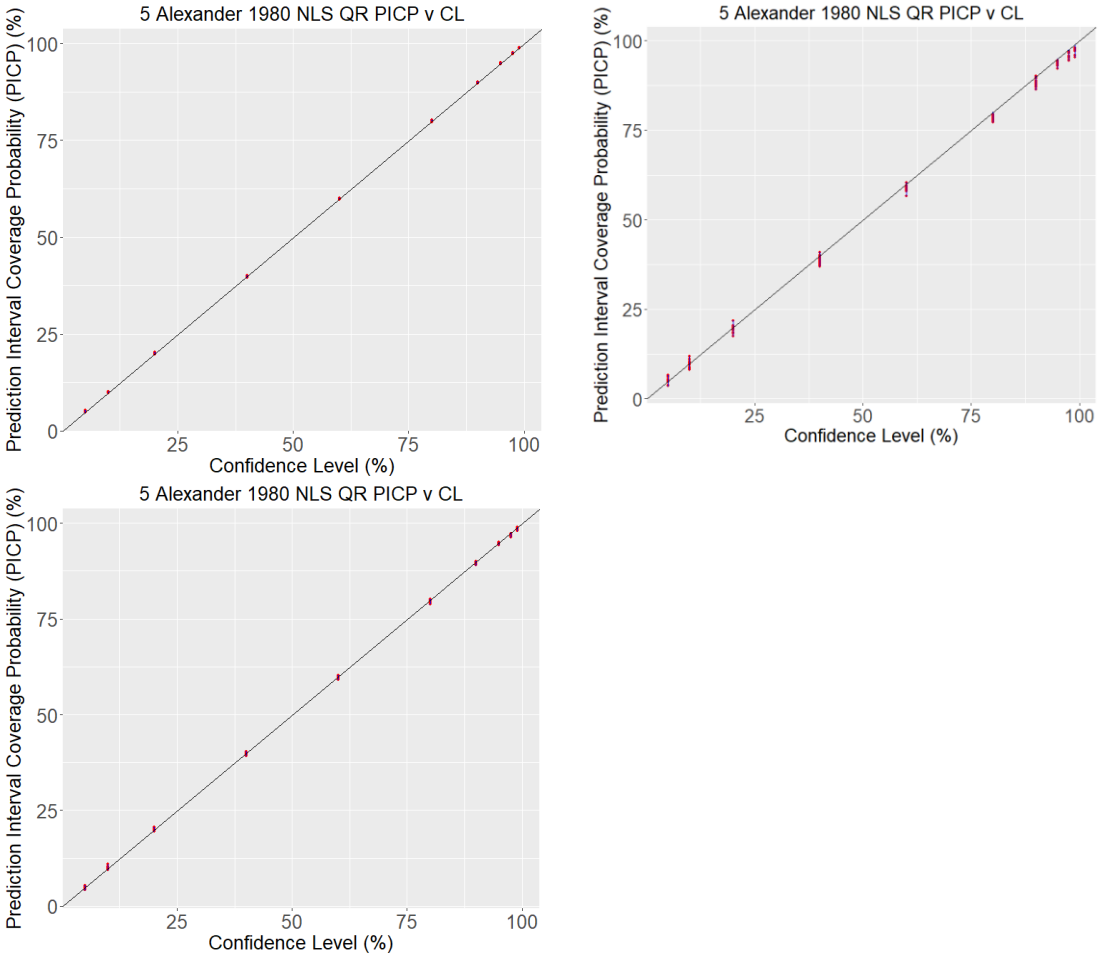
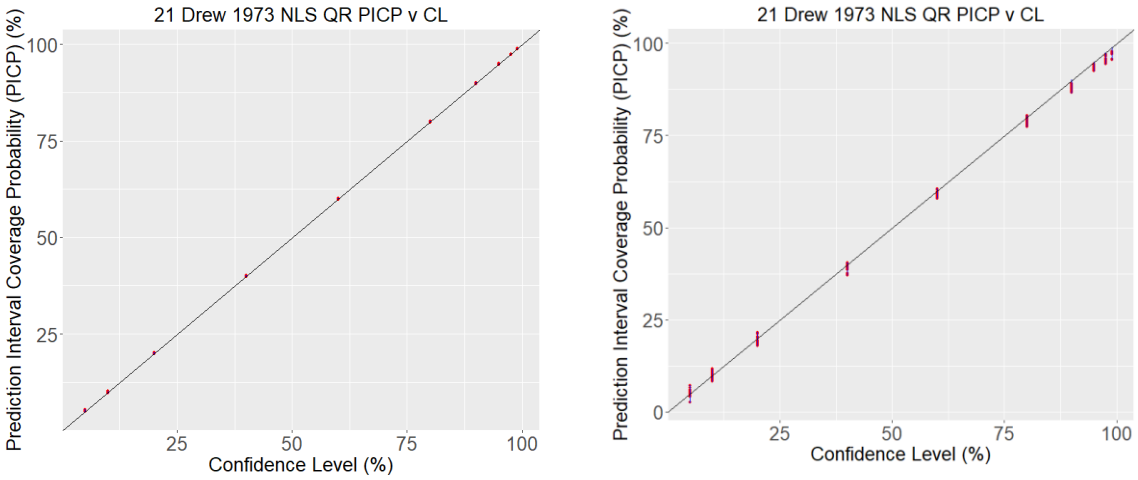


Figure 2.17. PICP vs CL Graphs for each dataset. For Model Group B (radical root functions), a representative PTF was chosen, and the PICP vs CL graph for that PTF is shown for each dataset. Ontario (All Variables) is shown top left; BC (All Variables) is shown top right; and BC (Carbon and Bulk Density) is shown bottom left.



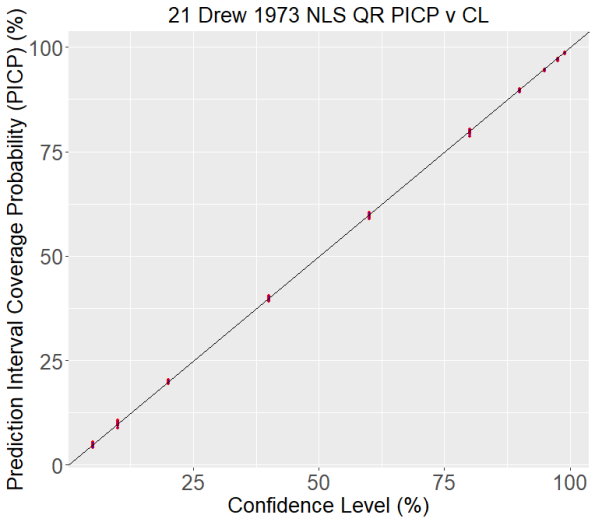
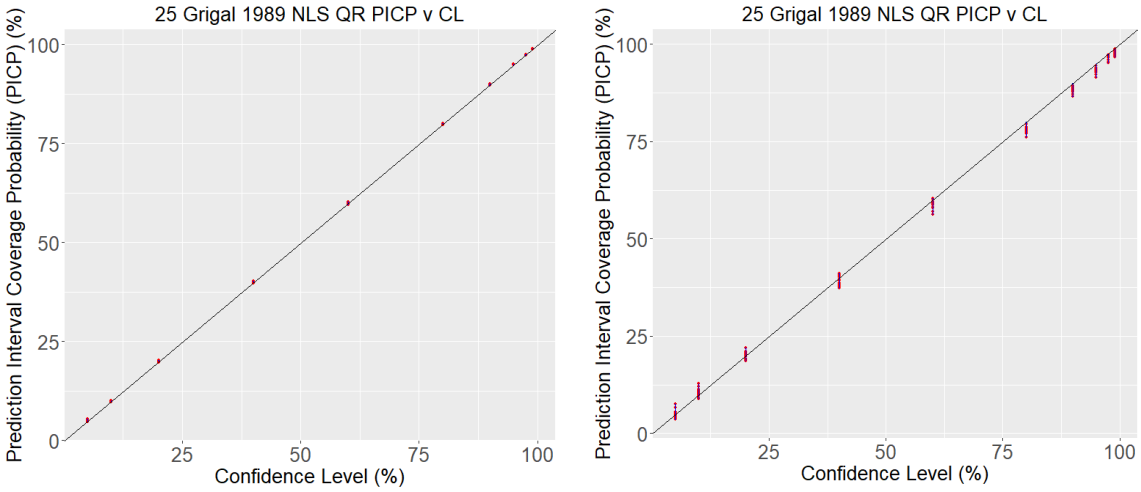


Figure 2.18. PICP vs CL Graphs for each dataset. For Model Group C (reciprocal functions), a representative PTF was chosen, and the PICP vs CL graph for that PTF is shown for each dataset. Ontario (All Variables) is shown top left; BC (All Variables) is shown top right; and BC (Carbon and Bulk Density) is shown bottom left.



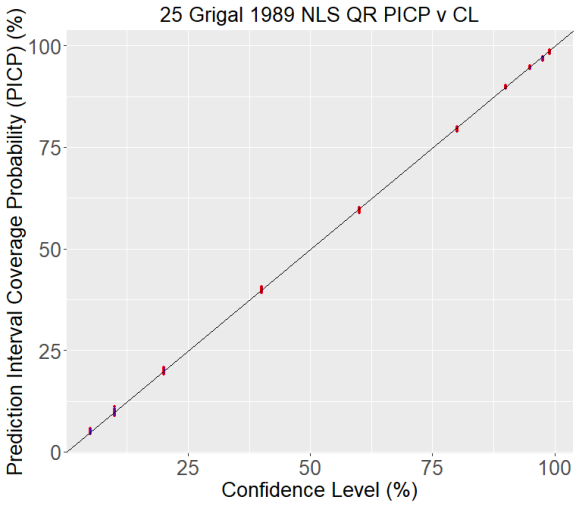


Figure 2.19. PICP vs CL Graphs for each dataset. For Model Group E (functions with natural exponent terms), a representative PTF was chosen, and the PICP vs CL graph for that PTF is shown for each dataset. Ontario (All Variables) is shown top left; BC (All Variables) is shown top right; and BC (Carbon and Bulk Density) is shown bottom left.

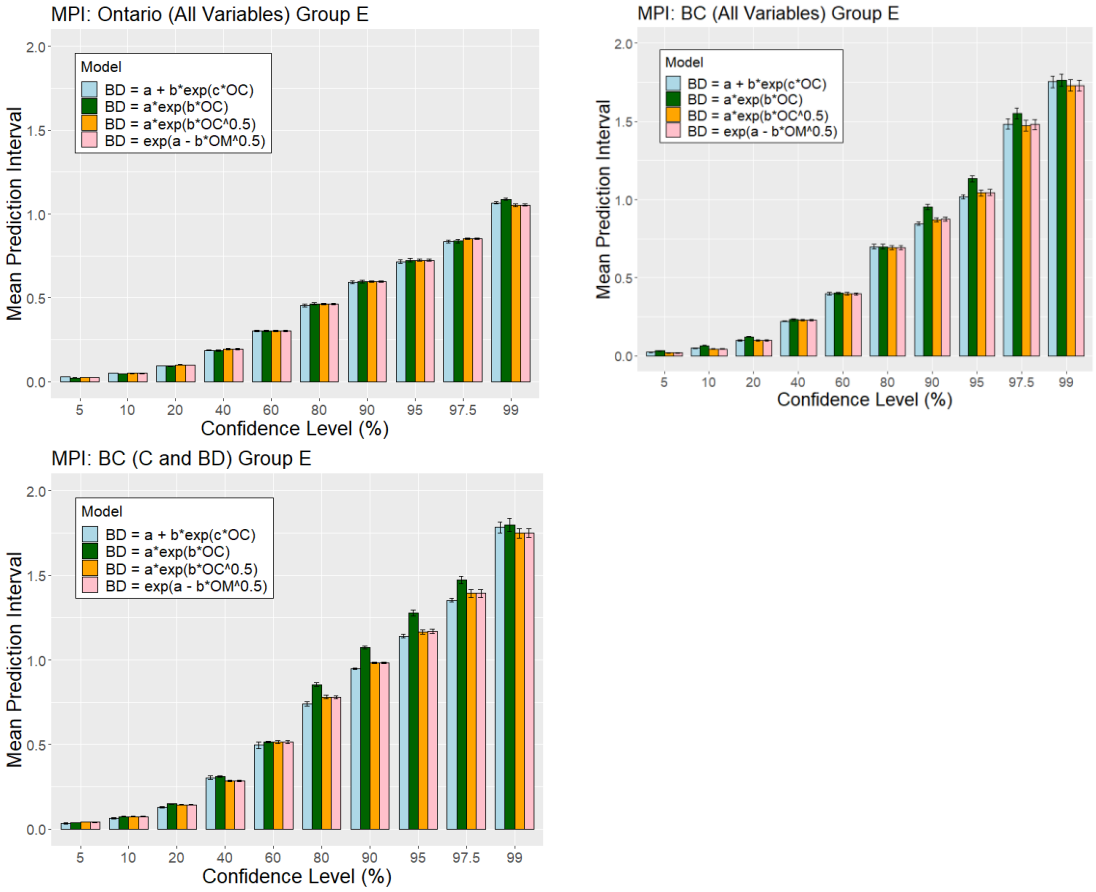


Figure 2.20. MPI Graphs for Model Group E PTFs, after recalibration with NLS, for each dataset. Ontario (All Variables) is shown top left; BC (All Variables) is shown top right; and BC (Carbon and Bulk Density) is shown bottom left.

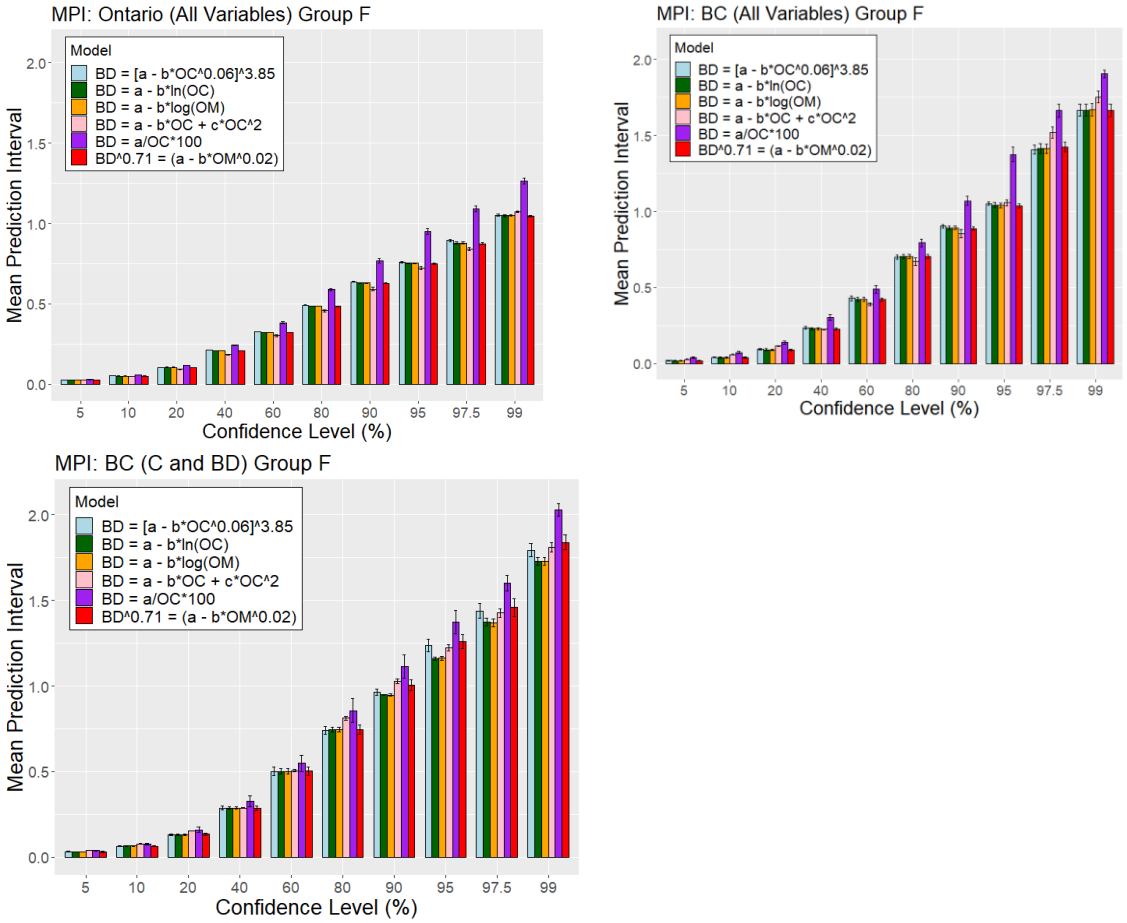


Figure 2.21. MPI Graphs for Model Group F PTFs, after recalibration with NLS. Ontario (All Variables) is shown top left; BC (All Variables) is shown top right; and BC (Carbon and Bulk Density) is shown bottom left.

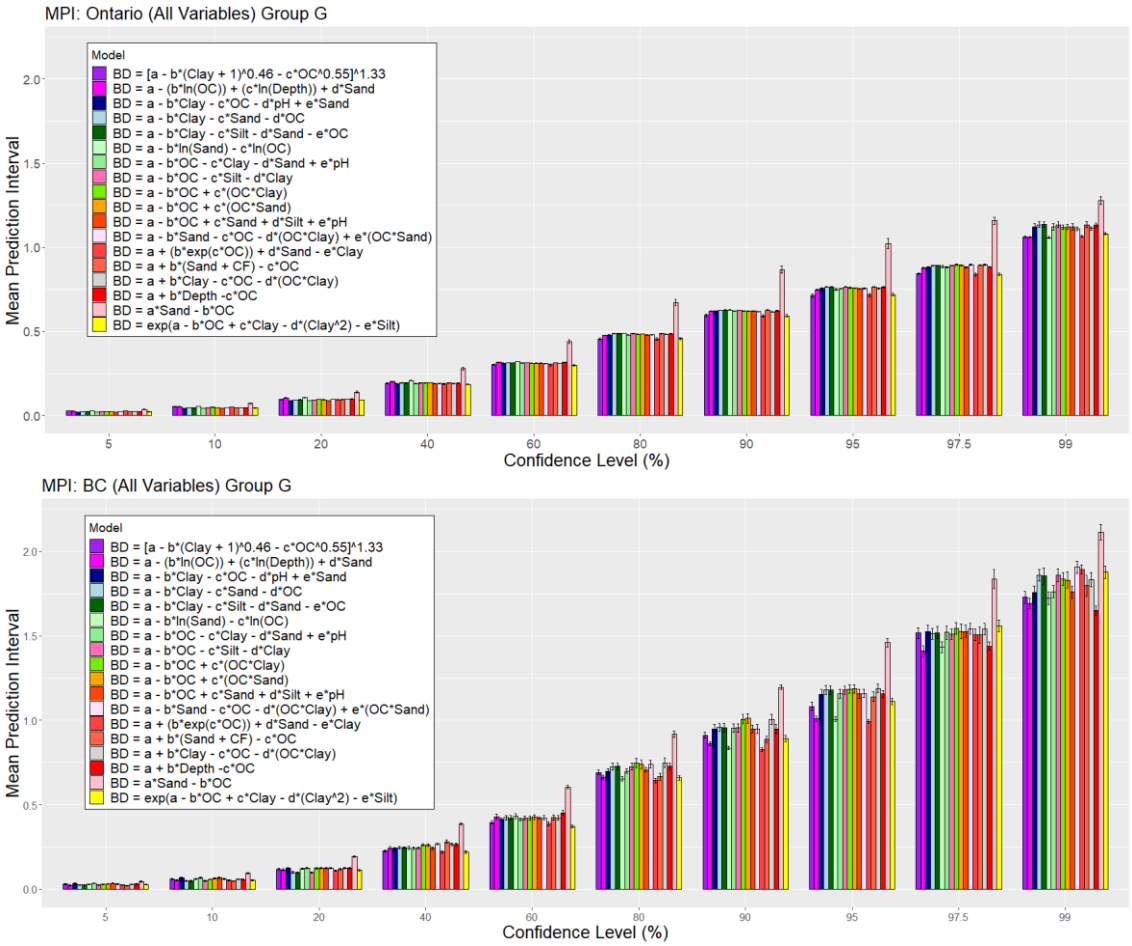


Figure 2.22. MPI Graphs for Model Group G PTFs, after recalibration with NLS. Ontario (All Variables) is shown top; BC (All Variables) is shown bottom.

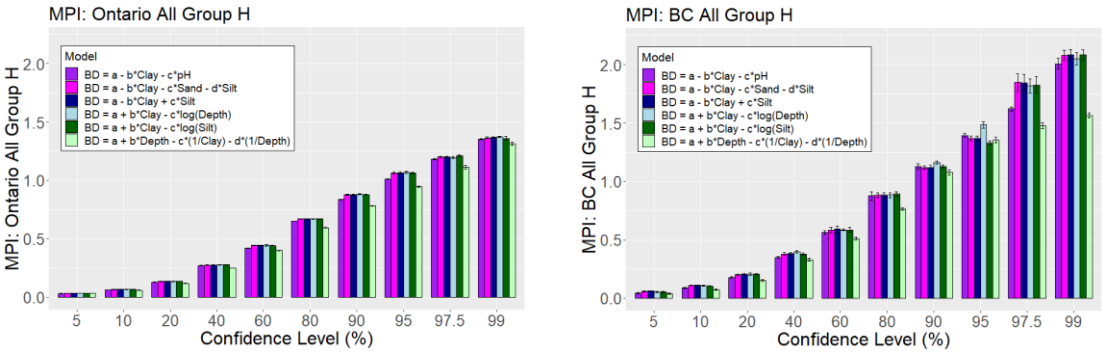


Figure 2.23. MPI Graphs for Model Group H PTFs, after recalibration with NLS. Ontario (All Variables) is shown left; BC (All Variables) is shown right.

Chapter 3. Machine Learning Approaches to Gap Filling Using Multiple Models Produced with Random Forest and Uncertainty Estimates Generated by Quantile Regression

3.1. Abstract

Legacy soil datasets are a valuable resource and should be used to the greatest extent possible. However, such datasets may be incomplete, and lack observations for every attribute, as the dataset may be a compilation of multiple studies. To use these datasets in soil mapping and modeling work, it is useful to fill the gaps in the dataset with estimated values. Machine learning is an approach that can provide estimates with high accuracy. In this study, the machine learner Random Forest (RF) was used to estimate bulk density values in an existing dataset from the province of British Columbia (BC), Canada, which was used as a case study dataset. As the dataset had missing observations across all attributes, multiple models need to be generated and tested to determine the accuracy of the estimated values produced. A total of 513 models were tested using RF, which were then ranked based on the concordance correlation coefficient (CCC) of their estimates; the CCC of all models ranged from 0.51 to 0.92. The estimates of 27 of these models were then used to fill the missing observations in the dataset; the accuracy of these models ranged from a CCC of 0.63 to 0.92. Further, uncertainty estimates for the predictions were generated using Quantile Regression (QR), which was coupled with the RF approach. Each model tested therefore had an accuracy measurement and an uncertainty estimate. This approach, of using multiple models developed in RF, can be applied to other legacy soil datasets with inconsistently missing values to produce estimates which can fill the missing observations, and produce uncertainty estimates for those estimates.

3.2. Introduction

Soil has many important functions: from an essential role in agriculture to being the largest reservoir of terrestrial carbon (FAO, 2004). Soil data are a valuable resource; however, it is often cost prohibitive to collect new data, so existing soil data must be used to the greatest extent possible (Arrouays et al., 2017). As soil datasets can be

compiled from many different sources, there are often gaps in the data, leading to incomplete coverage of soil attributes. Pedotransfer functions (PTFs) have been used to estimate values missing from soil databases; for example, bulk density is a frequently missing attribute, and there are many PTFs, which have been developed to estimate this property (Abdelbaki et al., 2018; Nasta et al., 2020; Nanko et al., 2014; Sevastas et al., 2018). In cases where the available data is patchwork, multiple models with different input variables must be applied to provide estimates for all horizons in the dataset. The estimates produced may vary in their accuracy and uncertainty, but their inclusion in soil datasets allows a greater amount of legacy data to be utilized. When used in conjunction with digital soil mapping, the estimates produced by PTFs increase the density of data available for mapping purposes.

Many regions have sparse soil data coverage, and there are not always financial resources available to acquire new data; however, using legacy soil data to the greatest extent possible is an option to provide data for DSM projects (Lagacherie, 2008). To discard existing data is “wasteful and scientifically irresponsible” (Rossiter, 2008). The density of coverage of existing data varies by country, and there is little traditional soil surveying currently taking place (Hartemink, 2008). Legacy data may come in the form of soil maps or soil profile data (Lagacherie, 2008); these data can be “renewed” so that it is accessible and useable (Rossiter, 2008). Legacy soil data have been used in digital soil mapping (DSM) projects in many regions around the world; Hendriks et al. (2019) found that regional studies relied more heavily on legacy soil data than did local studies. These legacy datasets are often augmented before modeling takes place using PTFs; for example, Silatsa et al. (2020) modelled soil carbon stock in Cameroon and used the Adams (1973) PTF to estimate missing BD values before calculating carbon stocks.

PTFs can take different forms: they may be in the form of look-up tables; regression-derived equations; or relationships developed through machine learning (Van Looy et al., 2017). The term “pedotransfer function” was introduced by Bouma (1989), who developed a PTF through regression analysis, which estimated bulk density from organic matter and particle size fraction. Earlier researchers had also developed equations relating soil attributes, such as Curtis and Post (1964) who developed an equation relating the density of pure mineral matter and pure organic matter to soil bulk density; or Breeuwsma et al. (1986) who estimated cation exchange capacity (CEC) from organic matter and clay. PTFs have been developed for specific soil conditions

and regions around the world, and there has been consensus that PTFs should not be applied to areas or soil conditions for which they were not developed (McBratney et al., 2002; De Vos et al., 2005; Benites et al., 2007; Casanova et al., 2016). Recalibrating an existing PTF for a regional dataset is an approach that improves the accuracy of the predictions generated by the PTF (Arbor et al., 2023). Recalibration can be performed using the nonlinear least squares method to generate new coefficients for existing PTFs (De Vos et al., 2005; Nanko et al., 2014; Khodaverdiloo et al., 2022; Arbor et al., 2023); the updated function is then better fitted to the dataset being used.

To generate a new PTF, an increasingly popular option is the use of machine learning (Padarian et al., 2020). An advantage of machine learners is that they are able to handle non-linear relationships (Padarian et al., 2020). While PTFs produced through machine learning do not produce an equation, they are based on a model with input variables whose importance may vary by region or soil conditions. As examples, Yi et al. (2016) used OC, silt, clay and depth as predictors for bulk density in China; Ramcharan et al. (2017) used OC, clay, sand, pH and depth for the United States; Chen et al. (2018) used OC, clay, sand, gravel, pH, silt and depth as input variables, while finding that OC and clay were the most important variables, for a region in France. Even PTFs developed in similar geographic areas may not produce accurate results when applied to different datasets; it has been asserted that it is the underlying data structure, and correlation between soil attributes, that determines the performance of a PTF (Fuentes-Guevara et al., 2022; Laurence et al., 2023). This supports the argument that a PTF should either be recalibrated for a given dataset, or a new PTF developed.

There are multiple different learners to choose from when producing PTFs, but a frequently used machine learner is Random Forest (Breiman, 2001). Random Forest has been used to estimate bulk density in many studies (Hikouei et al., 2021; Liu et al., 2021; Ramcharan et al., 2017; Sequeira et al., 2014; Martinelli and Gasser, 2022; Zihao et al., 2022). Szabó et al. (2021) also generated PTFs using Random Forest, as well as accompanying uncertainty estimates. The produced PTFs were for soil hydraulic properties where they tested 32 combinations of variables to assess their ability to predict different soil hydraulic properties. Random Forest is an ensemble of tree learners. For each tree, a dataset is selected from the training data with replacement, and trees are not dependent on previous trees (Liaw and Wiener, 2002). The output is the average result of all the trees. Advantages of Random Forest include that it resists

overfitting (Breiman, 2001); it can handle both categorical and continuous variables; and it is easy to use (Liaw and Wiener, 2002).

The generation of uncertainty estimates to accompany estimates of soil attributes has been noted as an issue to be addressed by multiple authors (Minasny and McBratney, 2002; Tranter et al., 2010; Malone et al., 2011; Van Looy et al., 2017), as estimates should be accompanied by some information on their precision (Koenker, 2017). Minasny and McBratney (2002) identified two types of uncertainty associated with PTFs: uncertainty associated with the input variables, and model uncertainty; and discussed methods to quantify this uncertainty. In PTFs developed for soil hydraulic properties, uncertainty has been addressed by multiple studies; examples include Deng et al. (2009); Kotlar et al. (2019); and Chirico et al. (2010). Tranter et al. (2010) used the Shreshtha and Solomantine (2006) method of fuzzy k-means clustering, as well as the fuzzy k-means with extragrades method, to assess the uncertainty of PTF estimates through prediction interval coverage probability (PICP).

To provide uncertainty estimates along with the output of the machine learning predictions, quantile regression (Koenker and Bassett, 1978) can be used. Quantile regression (QR) is an alternative to traditional regression analysis of the relationship between a response and predictor variable. Regression analysis such as ordinary least squares models a function of the mean of the response variable (Staffa et al., 2019). However, to understand the median of the distribution of the response variable, if it is skewed, or other characteristics, then modeling the conditional quantiles is informative (Hao and Naiman, 2007). It has broad applications, having been used in fields such as medical research (Beyerlein, 2014) and econometrics (Conyon and He, 2017). QR has previously been used in soil science, by Lombardo et al. (2018) to produce soil maps of OC; also by Arbor et al. (2023) to quantify uncertainty of recalibrated equation-based PTFs for bulk density. Kasraei et al. (2021) used quantile regression post processing with four machine learners, as well as Quantile Random Forest (QRF), to generate uncertainty maps for digital soil mapping and found that coupling QR with a machine learner produced stable results. Schmidinger and Heuvelink (2023) compared quantile regression post processing of Random Forest to four other methods of producing probabilistic predictions and found it to produce the best probabilistic predictions alongside QRF.

Many studies have addressed how to predict missing soil variables, but an issue that has not frequently been addressed is how to handle inconsistently, missing predictors. When only one predictor is missing or has low coverage, a PTF can be developed and applied for the whole dataset (Akpa et al., 2016). While there may be a model that produces predictions with the highest accuracy, the input variables used to generate that model may not always be available for all observations of the target variable. If the goal is to fill an entire dataset, then multiple models, based on various combinations of predictors will need to be applied. Using a similar approach, Benke et al. (2020) tested over 500 models to determine the best models for predicting electrical conductivity (EC) and organic carbon. They used a Generalized Linear Mixed Model with Residual Maximum Likelihood estimation, and validated using 5-fold cross validation.

The objectives of this paper were (1) to apply the machine learning algorithm Random Forest to estimate a target variable based on a set of input attributes; (2) to generate uncertainty measurements associated with those predictions; and (3) to develop a framework for gap-filling soil datasets with varying coverage of input variables, using multiple machine-learning generated models, with accompanying uncertainty estimates. Bulk density was selected as the case study variable for this paper; bulk density often has low coverage, due to the nature of the data collection, which can be costly, time-consuming, and labour intensive (Harrison and Bocock, 1981; Kaur et al., 2002; Botula et al., 2015; Sequeira et al., 2014). The province of British Columbia (BC), Canada was used as the case study region, where only 1.5% of available soil observations have measured values for bulk density.

3.3. Methodology

Using the province of British Columbia (BC), Canada as a case study region, and bulk density as a case study variable, the machine learner Random Forest was applied to all possible combinations of the input variables available to generate potential models to fill missing bulk density values. These models were cross-validated using k-fold cross validation. Further, for each estimated value, an uncertainty estimate was generated using QR.

3.3.1. Study Area

The study area was the province of BC, Canada. The western border of the province is the Pacific Ocean and the Alaskan panhandle; the eastern border runs from 48° 17’52.9” W to the northwest through the Rocky Mountains, then north to 60° 00’00” N. As a very diverse province, divisions into regions based on climate, geography, and ecology have been made. Using a combination of climate, physiography, vegetation, and soil, BC was classified into 14 biogeoclimatic (BEC) zones. Further divisions of these zones into subzones and variants are based on other factors such as elevation (Ministry of Forests, 2003). More broadly, BC was divided into five physiographic regions (Valentine et al., 1978): the Coast Mountains and Islands, the Interior Plateau, Columbia Mountains, and Southern Rockies, Northern and Central Plateaus and Mountains, and the Great Plains. The Coast Mountains and Islands are characterized by a marine climate, with moderate temperatures throughout the year, and high amounts of precipitation (Valentine et al., 1978). Common tree species of the coastal forest include western redcedar, Douglas-fir, western hemlock, amabilis fir and Sitka spruce (BC Ministry of Forests, 2003). The Interior Plateau receives much less precipitation than the coastal region, as it falls in the rainshadow of the coastal mountains. The continental climate of the region results in a 25°C temperature range and periods of aridity (Valentine et al., 1978). It is characterized by low relief landscapes formed from lava flows and deposited glacial drift (Meidinger and Pojar, 1991), with grasslands and forests dominated by lodgepole pine (BC Ministry of Forests, 2003). The Columbia Mountains and Southern Rockies have highly varied climatic conditions; mountain slopes receive high amounts of precipitation annually, from 1500 mm to 2000 mm, with half of this precipitation in the form of snow (Valentine et al., 1978). Glacial and fluvial deposits are frequent in valleys, with colluvium on steep slopes (Meidinger and Pojar, 1991). Common tree species include Englemann spruce and subalpine fir at higher elevations, and western redcedar and western hemlock on lower slopes (BC Ministry of Forests, 2022). The Northern and Central Plateaus and Mountains region is topographically diverse, extensively covered by glacial drift left after glacial retreat (Meidinger and Pojar, 1991). Precipitation is varied, from 400 mm annually in western valleys, up to 2000 mm annually in the eastern part of the region on the slopes of the Rocky Mountains. Summers are short with cool temperatures, and winters are cold; as a result, soils are often frozen from mid-autumn to April (Valentine et al., 1978). Tree species such as

spruce, aspen and pine are mixed with areas of alpine tundra and black spruce bogs (Ministry of Forests, 2003). Lastly, the Great Plains are found to the north and east of the Rocky Mountains, and have a similar climate to the Northern and Central Plateaus and Mountains regions; with slightly warmer summers and lower precipitation (Valentine et al., 1978).

3.3.2. Datasets

Data was compiled from multiple studies and the BC Soil Information System (BCSIS), for a total of 101,722 observations. Soil attributes included bulk density, depth, organic carbon (OC), cation exchange capacity (CEC), total nitrogen (TN), pH, coarse fragment (CF) percent, and sand, silt, and clay percentages. Categorical variables included textural class, parent material, and soil order. The dataset was restricted to mineral soil samples, so any observations with > 17% organic carbon were removed, as were any horizons which had an "Organic" designation for soil order. Bulk density values ≥ 2.65 g/cm^3 were removed, as 2.65 g/cm^3 is the commonly assumed value for particle density (Kroetsch and Wang, 2008); therefore, bulk density values equal to or greater than this were assumed to be erroneous. Any horizons with no depth measurements were also removed. Table 3.1 contains the summary statistics of the soil attributes for the whole dataset.

Continuous Predictors

Bulk density had been measured using four methods, including volumeter, saran (clod), excavation, and core (Blake, 1965). These values were available as "field state", which includes all material <7.5 cm in diameter (Quesnel and Suttie, 1983). Coarse Fragment (CF) values are available and reported as a percentage value. Depth was calculated as the midpoint value between the measured upper and lower horizon values, and is expressed in centimeters. Only observations with depth values which indicated mineral soil were included; forest floor observations were removed. Cation exchange capacity (CEC) values are reported in meq/100g. Total Nitrogen (TN) values are reported in percent. pH values are presented as pH in H_2O; pH values which were reported as pH in $CaCl_2$ in their source were converted, using an equation which was derived by regressing values of pH_{H2O} and pH_{CaCl2} from the dataset, with an R^2 value of 0.73. The equation is as follows:

$$pH_{(H2O)} = 0.9757 * pH_{(CaCl2)} + 0.7143 \quad (1)$$

Organic carbon (OC) was determined through three laboratory methods, LECO (Wang and Anderson, 1998); Walkley-Black (Walkley and Black, 1934); and loss-on-ignition (LOI) (CITE). OC values were converted to LECO: OC values measured through Walkley-Black were converted using the following equation:

$$OC_{(LECO)} = 1.47 * OC_{(Walkley\text{-}Black)} \quad (2)$$

OC values measured through LOI were converted to LECO using equation 3:

$$OC_{(LECO)} = 0.48 * OC_{(LOI)} - 0.003 \quad (3)$$

Sand, silt and clay were reported as percentage values, and so no conversion was necessary. Measurements were made through dry sieving, wet sieving, the hydrometer method, and the pipette method (Quesnel and Suttie, 1983). Sand and clay are measured values, while silt is calculated as 100 – (sand + clay), so silt was not included as a predictor.

Categorical Predictors

Soil order was used as the soil taxonomic classification. Horizons with the "Organic" order designation were removed, as the dataset was restricted to mineral soil horizons. Also, any Cryosol observations were removed due to low representation in the dataset. The three most common orders were Brunisolic, Podzolic, and Gleysolic.

Textural class was determined through hand texturing performed in the field. The textural classifications are those of the Canadian System of Soil Classification (Soil Classification Working Group, 1998). The most dominant classifications were Loamy Sand, with 12,591 observations, and Silty Loam with 12,376 observations. The textural classes with the lowest representation were Sandy Clay, with only 73 observations, and Silt, with 370 observations.

Initially, there were 20 classes of parent material. Observations with the classes Ice, Bog Morainal, and Anthropogenic were removed. Other classes were combined: Active Eolian, Eolian, and Glacial Eolian were all classified as Eolian; Active Fluvial and Fluvial together classified as Fluvial; Active Marine and Marine were combined to be the

Marine class; Colluvial and Inactive Colluvial were combined to make the Colluvial class; and Lacustrine and Glacial Lacustrine together were classified as Lacustrine. The other classes, Glacial Fluvial, Morainal, Organic, Volcanic, Saprolite, and Bedrock, were unchanged. The most common classes in the dataset were Marine, Morainal and Fluvial.

3.3.3. Model Development

The dataset contained 11 variables available for inclusion in developing predictive models for bulk density: depth, OC, CF, sand, clay, pH, TN, CEC, PM, textural class, and soil order. Measured values for these attributes were not evenly distributed throughout the dataset, and so to estimate a bulk density value for each horizon, multiple models needed to be employed. The first step was to determine all possible combinations of input variables; with 11 variables, this meant a calculated 2,037 number of combinations. Each combination was treated as a potential model to be tested; models ranged from one to eleven variables, containing every combination of variables possible. Depth has previously been found to be an important variable for bulk density estimation (Hengl et al., 2017), and depth is universally reported in the dataset (100% of horizons contain depth values). Therefore, it was decided to include soil depth as an input variable in every model, which reduced the number of potential models to be tested to 1,023. The variables sand, silt, and clay were almost always present in the dataset together, and together describe particle size distribution. Sand and clay are both measured variables, while silt is calculated as the difference between 100% - (sand + clay). As a calculated value, silt was therefore not included as an input variable. Sand and clay were always either both included, or both excluded in potential model combinations, as they showed high collinearity. The number of models was therefore reduced to 512.

When the number of potential models had been ascertained, then the training dataset size and number of potential horizons filled for each model were determined. Training dataset size was restricted by the number of bulk density horizons available, which at 1,450 represented only 1.43% of horizons in the dataset. The model with the largest training dataset was *BD = f(depth*, at 1,450 horizons; this model had the potential to fill 100,272 horizons with bulk density estimates. The model with the smallest training dataset was *BD = f(CEC + CF + depth + OC + order + pH + sand + clay + PM + textural class + TN)*, at 190 horizons; this combination included all 11 available attributes.

3.3.4. Accuracy Metrics

Models Tested using Random Forest

Every model was tested using the machine learner Random Forest (Breiman, 2001), which is available in the *caret* package for the R language (R Core Team, 2023). A feature of Random Forest is that it can handle both continuous and categorical data (Liaw and Wiener, 2002), which was important as both types of data were represented in the dataset. Due to the limited data available, and the need for applying multiple models to fill the gaps, cross-validation was used instead of partitioning the data into training and testing datasets. As soil measurements from the same profile can be correlated, a leave-profile-out cross-validation procedure was followed to reduce autocorrelation, and *k*-fold cross-validation was repeated 5 times.

Accuracy Assessment

The primary accuracy metric which was used to determine which model to apply was the concordance correlation coefficient (CCC). The agreement between the observed values and the predicted values is measured by CCC; the higher the CCC value, the greater the agreement. It is calculated using the following equation:

$$\mathrm{CCC} = \frac{2\rho\sigma_x\,\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \qquad (4)$$

where μ_x and μ_y are the means of the observed and predicted values and σ_x and σ_y are the variances of the observed and predicted values. The secondary metric used was root mean square error (RMSE). A lower RMSE value indicates a smaller average distance between the observed and predicted values, meaning that the model is well fitted to the data. It is calculated as:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{n}} \qquad (5)$$

Where x_i is the observed value, and $\hat{x}_i$ is the predicted value of the i^{th} observation; while n is the number of observations.

Uncertainty Estimates using Quantile Regression

Uncertainty estimates have not typically been supplied with newly developed PTFs (Van Looy et al., 2017; Deng et al., 2009), however this has been identified as an area for improvement (Minasny and McBratney, 2002). Quantile regression (QR) was chosen to produce uncertainty estimates to accompany each PTF developed. Its advantages include that it is not sensitive to outliers (Staffa et al., 2019) and provides information about the effect of the predictors on selected quantiles rather than just the mean (Das et al., 2019). It has previously been used to estimate uncertainty of PTFs produced using Random Forest (Rahmati et al., 2019); in digital soil mapping using multiple machine learners (Kasraei et al., 2021; Deragon et al., 2023; Schmidinger and Heuvelink, 2023); and has been coupled with a nonlinear least squares method of recalibrating PTFs (Arbor et al., 2023). A thorough explanation of QR and how it can be integrated into predictive modelling is available in Kasraei et al. (2021); the approach here follows the use of QR in previous studies (López López et al., 2014; Dogulu et al., 2015; Rahmati et al., 2019; Arbor et al., 2023).

In QR, observed values (y) and predicted values ($\hat{y}$) for each quantile (τ) are assumed to have a linear relationship:

$$y = a_\tau \hat{y} + b_\tau \quad (6)$$

where a_τ is the slope and b_τ is the intercept of the linear regression. To determine a_τ and b_τ, the sum of residuals is minimized in the following loss function:

$$min \sum_{j=1}^{J} \rho_\tau \left(y_j - \left(a_\tau \hat{y}_j + b_\tau\right)\right) \quad (7)$$

where y_j and $\hat{y}_j$ are the jth paired samples, J is the total number of samples, and ρ_τ is the QR function for the quantile τ:

$$\rho_\tau\left(\epsilon_j\right) = \begin{cases} (\tau - 1)\epsilon_j, & \epsilon_j < 0 \\ \tau\epsilon_j, & \epsilon_j > 0 \end{cases} \quad (8)$$

where ϵ_j are the model residuals, calculated as the difference between the observed and predicted values from Eq (7), for the τ^{th} quantile. For the desired quantile τ, the QR function is applied to the residual, ϵ_j in Eq (8).

The uncertainty can be expressed through the prediction interval confidence probability (PICP). The estimated prediction limits are calculated for a significance level of 1 – α, and the probability that the observed values fall within the prediction limits is the PICP (Dogulu et al., 2015). It is expressed in the following equation, where PL_t^{upper} is the upper prediction limit, PL_t^{lower} is the lower prediction limit, and y_t is the observed value:

$$\text{PICP} = \frac{1}{n}\sum_{t=1}^{n} C,\ C = \begin{cases} 1, PL_t^{lower} \leq y_t \leq PL_t^{upper} \\ 0, otherwise \end{cases} \quad (9)$$

The mean prediction interval (MPI) is calculated as follows:

$$\text{MPI} = \frac{1}{n}\sum_{\tau=1}^{n}(PL_\tau^{upper\ limit} - PL_\tau^{lower\ limit}) \quad (10)$$

as with PICP, PL_t^{upper} is the upper prediction limit and PL_t^{lower} is the lower prediction limit.

Gap Filling

Each model was tested, using the available horizons with the variables included in the model. As each model was different, so too were the number of horizons available for use in training the models. Along with the BD estimates produced by the model, its accuracy was measured using CCC, and models were then ranked based on the CCC value of their estimates. The model with the best CCC was applied first; subsequent models were then applied, until all data points had been filled. Each point therefore has an associated CCC and uncertainty. A diagram of the procedure is available in Figure 3.2.

3.4. Results and Discussion

In total, 512 models for BD estimation were tested. The models were ranked by CCC value, and the top 40 models are available in Table 3.3; the last 10 models are available in Table 3.4. All models used for gap filling, along with their rank, CCC value, number of horizons used in model training and number of horizons filled, are available in Table 3.5. The model with the highest CCC (0.92) was *BD = f(depth + CF + OC + pH + sand + clay)*, which filled 6,379 missing values. The model with the lowest CCC used was *BD = f(depth)*, with a CCC of 0.63, which estimated the remaining 528 bulk density values. This was not the worst performing model tested; the model with the lowest CCC

was *BD = f(depth + TN + order + textural class)*, with a CCC of 0.51, however this model was not used to estimate any missing values. The number of models for each CCC value range is shown in Figure 3.3. The mean CCC of all models tested was 0.82 and the median CCC was 0.85. The mean of models used in gap filling was 0.87, while the median of those models was 0.89.

The number of estimated horizons filled varied by model. Frequently, the BD value of a horizon would be estimated by multiple models, especially those horizons with many available attributes; therefore, the estimate with the highest CCC was used. The actual number of gaps filled by a given model was therefore usually less than the initial calculated potential number of horizons for which the model could provide estimates. As illustration of this, the second ranked model had the potential to fill 7,088 horizons, but 6,379 of these horizons had already been filled by the first ranked model, and so the second model filled only 709 horizons.

The difference in CCC value between models was sometimes extremely small; for example, there is only a 0.0025 difference between the first and second ranked model. Further, there are occurrences where several models have identical CCC values. When the values are identical, the model with the greatest number of horizons which could be estimated was chosen. This usually translated into the model with the fewest variables for a given CCC value being chosen; an advantage of a parsimonious model is that it is more likely to be used in other studies, if fewer predictors are required (Laurence et al., 2023). As an example, five models had a CCC value of 0.90, and were ranked as 30 to 34 of all models tested. Of these five, the model which had the fewest variables, and could fill the greatest number of horizons, was used.

3.4.1. Model Performance and Variables

The top 40 models had CCC values ranging from 0.92 for the first ranked model, to 0.90 for the 40th ranked model. The variables frequently appearing in these models include OC, sand and clay, pH, CEC, and CF. Order also was also often included in well performing models but was not as well represented in the models ultimately used for gap filling. OC as well as organic matter has been found to be highly correlated with BD and used in many PTFs (Abdelbaki, 2018; Adams, 1973; Alexander, 1980; Alexander, 1989; Curtis and Post, 1964; Drew, 1973; Federer et al., 1993; Grigal et al., 1989; Han et al.,

2012; Hossain et al., 2015; Kobal et al., 2011; Manrique and Jones, 1991; Song et al., 2005; Tamminen and Starr, 1994). OC appeared in 8 of the 27 models used in gap filling, and 29 of the top 40 ranked models. Likely the reason it did not appear in more of the models was due to its limited availability in the case study dataset; only 17% of horizons had an OC measurement.

Particle size fractions such as clay, silt and sand have also been found to be influential variables for the prediction of bulk density (Heinonen, 1977; Qiao et al., 2019; Tomasella and Hodnett, 1998). While sand and clay had very limited availability in the dataset, with only 9.8% of horizons having a measured value, they were highly represented in the top 40 models; second only to OC, sand and clay were present in 28 of the models, and in 10 of the models used for gap filling. The inclusion of OC with particle size fractions has also shown good results in published PTFs (Botula et al., 2015; Hollis et al., 2012; Kätterer et al., 2006; Kaur et al., 2002). In a study on the influence of OC on soil physical properties, Dexter et al. (2008) recalibrated a model with the form 1/BD = a + b(OC) + c(clay) using the Marquardt-Levenberg algorithm. They fit this model for two types of soil, one with high OC content, the other with low OC content. They found 1/BD to be highly correlated with OC content in the soil with low OC content; however, in the soil with high OC content, 1/BD was strongly correlated with the clay content rather than the OC content. They posit that this is due to complexed organic carbon (COC), which they defined as the association of 1g of OC with n grams of clay; their findings indicate that it is COC content that is more strongly correlated with BD rather than OC. Qin et al. (2022) found that in soils with low OC content in China, a simple linear model with clay as the only variable produced the most accurate PTF. In a study where both the OC and clay content are variable and not uniform, including both in a PTF would likely result in increased performance; in this study the best performing model included both OC and clay.

The inclusion of pH with particle size fractions or other variables has also yielded good PTFs in previous studies (Barros and Fearnside, 2015; Bernoux et al., 1998; Brahim et al., 2012; Pereira et al., 2016). pH was strongly represented in the top 40 models, appearing in 24 of them; further, it was the variable with the largest number of appearances in models used in gap filling, at 16 of 27 models. Among other factors, pH has been shown to be affected by agricultural tillage practices (Li et al., 2020), glaciation patterns (Balstrøm et al., 2013), geology (Kassai and István, 2018), and topography

(Zhang et al., 2022). Environmental attributes such as topography, vegetation, or land use were not included in the models, as these attributes will be used in further mapping projects which include the predicted BD values. However, pH may be acting as a proxy variable for some of these attributes, and this may be a contributing factor in its frequent appearance in high performing models.

The soil attributes which contribute the most to the prediction of CEC are clay content, OC, and pH (Minasny and Hartemink, 2011). As both OC and clay have been shown to be strong predictors for BD, it is unsurprising that a correlated attribute such as CEC would then also be a good predictor for BD. CEC was included in 21 of the top 40 models, and 7 of the models used in gap filling. Like many of the continuous variables used in this study, CEC had limited representation in the dataset, with only 14% of horizons containing a measured value. De Souza et al. (2016) included CEC as a predictor in models developed through MLR and RF; they found that models which used soil properties were more accurate in predicting BD than environmental variables; their RF model showed an R2 of 0.51.

CF was the third most prevalent attribute in the top 40 models, and was included in 22; it was also in 13 of the 27 models used in gap filling. CF is required for the calculation of fine fraction bulk density and soil organic carbon stocks (Mehler et al., 2014). CF, or gravel as it is also referred to as, was the second most important variable in Martin et al.'s (2009) PTF Model M, which produced an R2 of 0.94.

The last continuous variable included in the models was TN; it was present in 16 of the top 40 models, and 5 of the models used in gap filling. TN has previously been found to be a strong predictor of BD, and highly correlated with OC, such that it could be used as a replacement for OC in PTFs (Benites et al., 2007). Han et al. (2012) also found TN to be an important predictor for BD.

Variables with high numbers of observations in the dataset included order (90.06%), PM (93.56%), CF (98.42%); and depth (100%). Textural class was the categorical variable with the lowest coverage, at 84.40%. The models with the poorest performance among all models tested were dominated by categorical input variables, such as textural class, PM, and order. The advantage of the inclusion of these variables, however, is that they were well represented in the dataset. Models with lower CCC

values but with high coverage were then used to estimate those missing values without other variables. The model which filled the largest number of horizons was *BD = f(depth + CF + order + PM + textural class)*, which was the 25th model applied, but the 392nd ranked model overall. This model contained only categorical variables plus depth, but filled almost 64% of horizons with missing values. Relative to other models tested, the CCC value was moderate, but still near the mean of models used in gap filling, at 0.86.

Soil order was the only categorical variable represented in the top 40 models, appearing in 7, and 3 of the models used in gap filling. Soil order is a Canadian taxonomic classification, but soil classifications have been used in previous PTFs for predicting bulk density. Soil taxon was included in Martin et al.'s (2007) PTF developed through Multiple Additive Regression Tree modelling, and was ranked the sixth most important predictor. Rather than use soil classifications as input variables, several studies have grouped data and developed PTFs for the classification. Heuscher et al. (2005) found that the performance of PTFs improved when data was divided into soil suborders for a US national dataset; Manrique and Jones (1991) similarly found that this method improved the results for some suborders. Other studies, such as de Souza et al. (2016) did not find any benefit to grouping data by soil classification. With the very limited availability of BD data in this study, subdividing the data was not an option, and so soil order was used as an input variable.

Neither parent material nor textural class were represented in the top 40 models; however, 3 of the models for gap filling included parent material, and 1 included textural class, which was the model that predicted the majority of the missing BD values. Calhoun et al. (2001) developed PTFs which included both texture and parent material, and found that combining continuous, lab-derived variables with categorical, site descriptive variables produced the best models. They also investigated grouping their data by parent material type, and found it to improve the accuracy of the estimations.

3.4.2. Comparison of Results to Other Studies

In a study with a similar approach to estimating missing values, Benke et al. (2020) filled gaps in an existing dataset, with data gathered from almost a hundred different projects over 66 years. As in the dataset used in this study, many samples did not have the full suite of variables available, and so they tested 560 models to predict

each of EC and OC. Model performance was evaluated using mean squared prediction error (MSPE) and mean absolute percentage error (MAPE), and then the models were ranked based on their MSPE. The authors found a tradeoff between the number of potential filled values vs model performance; the 1st ranked model used all available variables and had an MSPE of 0.686, while the 35th ranked model for EC had an MSPE of 0.710, but could predict a large amount of missing values. This is similar to the results produced by this study, where the highest ranked model had a high CCC and filled a sizeable number of horizons at 6,379; however, the model which filled 63.9% of the horizons in the dataset was ranked 392nd. Unlike Benke's finding, the highest ranked model in this study did not contain all available variables; it only contained 6 of 11 potential variables. The model which used all available variables was ranked number 225, with a CCC of 0.85, which was greater than the median value of all models tested.

Comparison to other studies can be hampered by the metrics used to evaluate the accuracy of the PTF; for this study, CCC was chosen as the method of ranking the PTFs performance. Akpa et al. (2016) also used CCC to assess the PTFs which they developed. Multiple linear regression (MLR) and RF were used to estimate bulk density, incorporating environmental variables as well as soil attributes as input. They compared using only soil attributes or only environmental data with using both soil and environmental data, and found that the latter combination yielded the highest performing PTFs. Using RF, the highest CCC values were 0.800 for all data; when the samples were divided into topsoil and subsoil, the CCC results were 0.608 for topsoil (<0.40 m) and 0.839 for subsoil (>0.40 m). RMSE values for all data were 0.107 Mg/m3, 0.118 for topsoil and 0.102 for subsoil. This compares to the most accurate PTF produced in this study, with a CCC of 0.92 and an RMSE of 0.10 g/cm^3. Due to limited data availability, this study did not subdivide the dataset based on depth or other characteristic, although other studies have found success with developing PTFs for specific designations such as horizon (Reidy et al., 2016) when filling missing values in a dataset.

A previously used approach of recalibrating existing PTFs improved the accuracy of those PTFs (Arbor et al., 2023). The recalibrated PTFs were equation-based functions for the prediction of BD developed through regression analysis. After recalibration almost all PTFs showed improvement, with the highest reported CCC value of 0.68 for a PTF which included OC, clay and silt as predictors for BD. Overall, the study found that fewer predictors and simpler model forms were more easily recalibrated and produced the

most accurate estimates. In comparison, accuracy did not decrease with the addition of more predictors, and often increased. Further, the machine learner RF allowed easy incorporation of categorical variables into the models, while categorical variables were not included in any of the equation-based functions. Finally, the CCC values of the highest performing model in this study was significantly higher than the CCC values of the recalibrated PTFs.

Palladino et al. (2022) used soil attribute variables and environmental variables to create eleven PTFs, which contained different combinations and numbers of variables, using RF. The best performing PTF had a Pearson correlation coefficient of 0.616, and RMSE value of 0.163 g/cm^3. This PTF used sand, clay, OC, pH, CaCO3, elevation, slope, and rock fragment. While many PTFs now incorporate environmental variables (Akpa et al., 2016; de Souza et al., 2016; Schillaci et al., 2021) this study focused on using only soil attributes as variables to fill missing gaps in the dataset; when further digital soil mapping is conducted using this augmented dataset, environmental variables will be incorporated in model development.

3.4.3. Training Dataset Size and Model Performance

The model used in gap filling with the largest training dataset was *BD = f(depth)*, which was the 27th and last model applied, and which used 1,450 horizons to train the model. The model with the smallest training dataset incorporated all 11 available variables and had a training dataset size of 191. With limited data availability, it is desirable to maximize the use of that data, and to determine the minimal sample size acceptable for modelling purposes. It can be difficult to decide what the lower limit for the size of the training dataset should be. Heuscher et al. (2005) addressed this topic, determining how many samples would be required for a valid MLR model. Based on Muller and Fetterman's (2002) work, they multiplied the number of predictor variables, plus one for the intercept, by 10 to calculate the minimum number of samples. Using this method, the minimum number of samples required for a model in this study with all 11 available variables and developed using MLR would be 120. With this assessment, the 191 horizons with the maximum combination of variables available in the case study dataset would be suitable to use for an MLR model.

There are multiple studies in the literature which investigate the use of different sample sizes for training machine learning algorithms. In a study which examined groundwater potential, Moghaddam et al. (2020) tested the effect of differing sample sizes on four machine learning algorithms, including RF. They found RF to perform the best of the four across all sample sizes. Wu et al. (2022) also used RF when investigating sample size effect on soil OC prediction. They used 10 sample sizes to train their model and found that the prediction accuracy increased as sample size increased, until the second to last sample size. They also noted that variable importance changed with sample size.

Somarathna et al. (2017) tested multiple machine learners on different sized training datasets and found that the accuracy of the estimations of soil carbon was related to the size of the dataset. As training dataset size increased, so did the accuracy of the estimations, until a levelling off point was reached, which varied by model. While there were differences in accuracy between the models tested, sample size had a greater effect on accuracy. They also noted that the uncertainty of the estimations was also affected by sample size and decreased as sample size increased. Our findings indicated that small training datasets could produce estimates with high accuracy, but also higher uncertainty than when larger training datasets were used.

3.4.4. Uncertainty Estimations

For each model used, two types of graphs were generated: a PICP vs CI graph, and an observed vs predicted plot. Examples of the observed vs predicted plot for the first and last models used to estimate missing values, are shown in Figure 3.4. For the observed vs predicted plots, the blue lines on the graph are the 95th and 5th quantiles, with the distance between the lines at any given point on the x axis being the 90% prediction interval (PI), which is the interval in which 90% of predicted values are expected to fall. As shown in Figure 3.4, the highest ranked model has a narrow PI width, indicating low uncertainty of future predictions falling within this interval. In contrast, the last model used has a very wide PI, meaning that the range of values in which future predictions could be expected is much greater, reflecting the higher uncertainty of the results.

Figure 3.5 presents the PICP vs CL graphs for the first and last used models; graphs for all other models used for gap filling are presented as Figures 3.7 to 3.31. PICP is the preferred metric to evaluate uncertainty over MPI (Dogulu et al., 2015). When examining the PICP vs CL graphs, for each confidence level, the PICP should be the same or similar (Dogulu et al., 2015); this means that values should be close to the 1:1 line. If the PICP is greater than its corresponding confidence level and shows a greater spread around the 1:1 line, it indicates greater uncertainty for the predicted values. Model 1 has a greater spread around the 1:1 line than Model 27; this also corresponds to the smaller training dataset size of Model 1 (381 horizons) than that of Model 27 (1450 horizons). Somarathna et al. (2017) found an inverse relationship between sample size and uncertainty; as sample size increased, uncertainty decreased. The same pattern was also observed by Arbor et al. (2023), where the PICP values reflected the uncertainty associated with the training dataset size.

A graph of the MPI values for a selection of the models used in gap filling is available in Figure 3.6. A smaller MPI value indicates lower uncertainty (Muthusamy et al., 2016); it is the average range in which a future estimation is expected to occur (Rahmati et al., 2019). MPI values are plotted for the models at selected confidence levels, ranging from 5 to 99%. A confidence level expresses the probability of a value occurring; at a 99% confidence level there is a 99% probability that a future estimate will occur within the MPI. For models with wide MPI values, it indicates that there is a large range of future expected estimates. Model 27 has much larger MPI values for each confidence level (CL) than for the other models used in gap filling, while Models 1, 5, and 10 have much lower MPI values for every confidence level. This reflects the differences in accuracy of estimations for each model.

3.5. Conclusion

To address the need to maximize the use of existing soil data, an approach to gap filling a legacy soil dataset with inconsistently missing values was applied. Using the available variables, every combination of those variables was identified and used as a predictive model to estimate missing values. Over 500 models were tested using Random Forest and ranked based on the CCC value of their estimations. The accuracy of the models ranged from a CCC of 0.92 for the first ranked model, to 0.51 for the last ranked model. Of these models, 27 were used to estimate bulk density values missing in

the dataset. The accuracy of the models used was 0.92 for the first model applied to 0.63 for the 27th model applied. Variables that were included in high accuracy models were depth, OC, CEC, CF, pH, sand, and clay; however, these variables had low representation in the dataset and so could not be used to predict all missing values. Models with lower accuracy but which could predict a larger number of variables were used to fill the majority of the missing values; variables in these models included soil order, textural class, and parent material. Through QR, uncertainty estimates of the models output were produced, and expressed through PICP vs CL graphs, and MPI plots. The PICP graphs showed minimal variation from the 1:1 line for the models used in gap filling, indicating the low uncertainty of these models; those models with the largest training datasets corresponded with the lowest uncertainty. The MPI graphs were used as a secondary uncertainty measure, and provided a range of values in which future predictions will fall for a given confidence level, and indicated that models with higher accuracy had lower MPI values. The approach presented here can be used in further studies for gap filling datasets with inconsistently missing data, and could be applied to other target variables or utilized with different machine learners.

3.6. References

Abdelbaki, A.M. 2018. Evaluation of pedotransfer functions for predicting soil bulk density for U.S. soils. Ain Shams Engineering Journal, **9:** 1611-1619.

Adams, W.A. 1973. The effect of organic matter on the bulk and true densities of some uncultivated podzolic soils. Journal of Soil Science, **24(1):** 10-17.

Aertsen, W., Kint, V., van Orshoven, J., Özkan, K., Muys, B. 2010. Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. Ecological Modelling, **221(8):** 1119-1130.

Akpa, S.I.C., Ugbaje, S.U., Bishop, T.F.A. and Odeh, I.O.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:** 644-658.

Alaboz, P., Demir, S., Dengiz, O. 2021. Assessment of various pedotransfer functions for the prediction of the dry bulk density of cultivated soils in a semiarid environment. Communications in Soil Science and Plant Analysis, **52(7):** 724-742.

Alexander, E.B. 1980. Bulk densities of California soils in relation to other soil properties. Soil Sci. Soc. Am. J., **44(4):** 689-692.

Alexander, E.B. 1989. Bulk density equations for southern Alaska soils. Can. J. Soil Sci., **69(1):** 177-180.

Arbor, A., Schmidt, M., Saurette, D., Zhang, J., Bulmer, C., Filatow, D., Kasraei, B., Smukler, S., Heung, B. 2023. A framework for recalibrating pedotransfer functions using nonlinear least squares and estimating uncertainty using quantile regression. Geoderma, **439:** 116674.

Balstrøm, T., Breuning-Madsen, H., Krüger, J., Jemsen, N.H., Greve, M.H. A statistically based mapping of the influence of geology and land use on soil pH: A case study from Denmark. Geoderma, **192:** 453-462.

Barros, H.S., Fearnside, P.M., 2015. Pedo-transfer functions for estimating soil bulk density in central Amazonia. R. Bras. Ci. Solo, **39:** 397–407.

Benites, V.M., Machado, P.L.O.A., Fidalgo, E.C.C., Coelho, M.R. and Madari, B.E. 2007. Pedotransfer functions for estimating soil bulk density from existing soil survey reports in Brazil. Geoderma, **139:** 90-97.

Benke, K.K., Norng, S., Robinson, N.J., Chia, K., Rees, D.B., Hopley, J. 2020. Development of pedotransfer functions by machine learning for prediction of soil electrical conductivity and organic carbon content. Geoderma, **366:** 114210.

Bernoux, M., Arrouays, D., Cerri, C., Volkoff, B., Jolivet, C., 1998. Bulk Densities of Brazilian Amazon soils related to other soil properties. Soil Sci. Soc. Am. J., **62:** 743–749.

Beyerlein, A., 2014. Quantile regression – opportunities and challenges from a user's perspective. American journal of epidemiology, **180(3):** 330-331.

Blake, G.R. Ch. 30, Bulk Density. In Methods of Soil Analysis, Part 1. 1965. C.A. Black, Editor-in-Chief, and D.D. Evans [and Others] Associate Editors; R.C. Dinauer, Managing Editor. American Society of Agronomy. 374-390.

Botula, Y-D., Nemes, A., Van Ranst, E., Mafuka, P., De Pue, J., and Cornelis, W.M. 2015. Hierarchical pedotransfer functions to predict bulk density of highly weathered soils in Central Africa. Soil Sci. Soc. Am. J. **79:** 476-486.

Bouma, J. 1989. Using soil survey data for quantitative land evaluation. Advances in Soil Science, **9:** 177-213.

Brahim, N., Bernoux, M., Gallali, T., 2012. Pedotransfer functions to estimate soil bulk density for Northern Africa: Tunisia case. J. Arid Environ., **81:** 77–83.

Breeuwsma, A., Wösten, J.H.M., Vleeshouwer, J.J., van Slobbe, A.M., and Bouma, J. 1986. Derivation of land qualities to assess environmental problems from soil surveys. Soil Sci. Soc. Am. J., **50:** 186–190.

Breiman, L. 2001. Random Forests. Machine Learning, **45:** 5-32.

Calhoun, F.G., Smeck, N.E., Slater, B.L., Bigham, J.M., and Hall, G.F. 2001. Predicting bulk density of Ohio soils from morphology, genetic principles, and laboratory characterization data. Soil Sci. Soc. Am. J., **65:** 811-819.

Casanova, M., Tapia, E., Seguel, O., Salazar, O. 2016. Direct measurement and prediction of bulk density on alluvial soils of central Chile. Chilean Journal of Agricultural Research, **76(1):** 105-113.

Chen, S.C., Richer-de-Forges, A.C., Saby, N.P.A., Martin, M.P., Walter, C., Arrouays, D. Building a pedotransfer function for soil bulk density on regional dataset and testing its validity over a larger area. Geoderma, **312:** 52-63.

Chirico, G.B., Medina, H., Romano, N. 2010. Functional evaluation of PTF prediction uncertainty: An application at hillslope scale. Geoderma, **155:** 193-202.

Conyon, M.J., He, L. 2017. Firm performance and boardroom gender diversity: A quantile regression approach. Journal of Business Research, **79:** 198-211.

Curtis, R.O. and Post, B.W. 1964. Estimating bulk density from organic-matter content in some Vermont forest soils. Soil Sci. Soc. Am. Proc. **28:** 285-286.

Das, K., Krzywinski, M., Altman, N. 2019. Quantile regression. Nature Methods, **16:** 451-452.

Deng, H., Ye, M., Schaap, M.G., Khaleel, R. 2009. Quantification of uncertainty in pedotransfer function-based parameter estimation for unsaturated flow modeling. Water Resources Research, **45:** W04409.

Deragon, R., Heung, B., Lefebvre, N., John, K., Cambouris, A.N., Caron, J. 2023. Improving a regional peat thickness map using soil apparent electrical conductivity measurements at the field-scale. Frontiers in Soil Science, **3:** 1305105.

De Souza, E., Fernandes Filho, E.I., Schaefer, C.E.G.R., Batjes, N.H., dos Santos, G.R., Pontes, L.M. 2016. Pedotransfer functions to estimate bulk density from soil properties and environmental covariates: Rio Doce basin. Scientia Agricola, **73(6):** 525-534.

De Vos, B., Van Meirvenne, M., Quataert, P., Deckers, J., and Muys, B. 2005. Predictive quality of pedotransfer functions for estimating bulk density of forest soils. Soil Sci. Soc. Am. J., **69:** 500-510.

Dexter, A.R., Richard, G., Arrouays, D., Czyż, E.A., Jolivet, C., Duval, O. 2008. Complexed organic matter controls soil physical properties. Geoderma, **144:** 620-627.

Dogulu, N., López López, P., Solomatine, D.P., Weerts, A.H., Shrestha, D.L. 2015. Estimation of predictive hydrologic uncertainty using the quantile regression and UNEEC methods and their comparison on contrasting catchments. Hydrology and Earth System Sciences, **19:** 3181-3201.

Drew, L.A. 1973. Bulk density estimation based on organic matter content of some Minnesota soils. Minnesota Forestry Research Notes, 243. Sci. Jour. Ser. Paper No. 8333 of the University of Minnesota Agricultural Experimental Station.

Federer, C.A., Turcotte, D.E., Smith, C.T. 1993. The organic fraction – bulk density relationship and the expression of nutrient content in forest soils. Can. J. For. Res., **23(6):** 1026-1032.

Food and Agriculture Organization (FAO) of the United Nations. 2004. Carbon sequestration in dryland soils. World Soil Resource Reports 102. Rome, Italy.

Fuentes-Guevara, M.D., Armindo, R.A., Timm, L.C., and Nemes, A. 2022. Data correlation structure controls pedotransfer function performance. Journal of Hydrology, **614:**128540.

Grigal, D.F., Brovold, S.L., Nord, W.S., Ohmann, L.F., 1989. Bulk density of surface soils and peat in the north central United States. Can. J. Soil Sci., **69(4):** 895–900.

Guo, L., Fan, G., Zhang, Y., and Shen, Z. 2019. Estimating the bulk density in 0-20 cm of tilled soils in China's Loess Plateau using support vector machine modeling. Communications in Soil Science and Plant Analysis, **50(14):** 1753-1763.

Gunarathna, M.H.J.P., Sakai, K., Nakadakari, T., Momii, K., and Kumari, M.K.N. 2019. Machine learning approaches to develop pedotransfer functions for tropical Sri Lankan soils. Water, **11:** 1940.

Han, G.-Z., Zhang, G.-L., Gong, Z.-T., Wang, G.-F., 2012. Pedotransfer functions for estimating soil bulk density in China. Soil Sci., **177(3):** 158–164.

Hao, L., Naiman, D.Q. 2007. Quantile Regression. Los Angeles, Calif; London: SAGE.

Harrison, A.F., and Bocock, K.L. 1981. Estimation of soil bulk-density from loss-on-ignition values. Journal of Applied Ecology, **8:** 919-927.

Hartemink, A.E. 2008. Soil map density and a nation's wealth and income. In: Hartemink, A.E., McBratney, A., Mendonça-Santos, M.L. (Eds). Digital Soil Mapping with Limited Data. Dordrecht; London: Springer.

Heinonen, R., 1977. Towards "normal" soil bulk density. Soil Sci. Soc. Am. J. **41(6):** 1214–1215.

Hendriks, C.M.J., Stoorvogel, J.J., Lutz, F., Claessens, L. 2019. When can legacy soil data be used, and when should new data be collected instead? Geoderma, **348:** 181-188.

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B.M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotic, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B. 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS ONE, **12(2):** e0169748.

Heuscher, S.A., Brandt, C.C., Jardine, P.M. 2005. Using soil physical and chemical properties to estimate bulk density. Soil Sci. Soc. Am. J., **69:** 51-56.

Hikouei, I.S., Kim, S.S., Mishra, D.R. 2021. Machine-Learning classification of soil bulk density in salt marsh environments. Sensors, **21:** 4408.

Hollis, J.M., Hannam, J., Bellamy, P.H., 2012. Empirically-derived pedotransfer functions for predicting bulk density in European soils. Eur. J. Soil Sci. **63:** 96–109.

Hossain, M.F., Chen, W., Zhang, Y., 2015. Bulk density of mineral and organic soils in the Canada's arctic and sub-arctic. Information Processing in Agriculture **2(3-4):** 183–190.

Kasraei, B., Heung, B., Saurette, D.D., Schmidt, M.G., Bulmer, C.E., and Bethel, W. 2021. Quantile regression as a generic approach for estimating uncertainty of digital soil maps produced from machine-learning. Environmental Modelling and Software, **144:** 105139.

Kassai, P. and István, S. 2018. The role of geology in the spatial prediction of soil properties in the watershed of Lake Balaton, Hungary. Geologica Croatica, **71(1):** 29-39.

Kätterer, T., Andrén, O., Jansson, P.-E., 2006. Pedotransfer functions for estimating plant available water and bulk density in Swedish agricultural soils. Acta Agriculturae Scandinavica Section B-Soil and Plant Science, **56(4):** 263–276.

Kaur, R., Kumar, S., and Gurung, H.P. 2002. A pedo-transfer function (PTF) for estimating soil bulk density from basic soil data and its comparison with existing PTFs. Aust. J. Soil Res., **40:** 847-857.

Khodaverdiloo, H., Bahrami, A., Rahmati, M., Vereecken, H., Miryaghoubzadeh, M., Thompson, S. 2022. Recalibration of existing pedotransfer functions to estimate soil bulk density at a regional scale. Eur. J. Soil Science, **73:** e13244.

Kobal, M., Urbančič, M., Potočič, N., De Vos, B., Simončič, 2011. Pedotransfer functions for bulk density estimation of forest soils. J. Forestry Soc. Croatia, **135:** 19–27.

Koenker, R. and Bassett Jr., G. 1978. Regression Quantiles. Econometrica, **46(1):** 33-50.

Koenker, R. 2017. Quantile regression: 40 years on. Annual Review of Economics, **9:** 155-176.

Kotlar, A.M., de Jong van Lier, Q., Barros, A.H.C., Iversen, B.V., and Vereecken, H. 2019. Development and uncertainty assessment of pedotransfer functions for predicting water contents at specific pressure heads. Vadose Zone Journal, **18:** 190063.

Kroetsch, D. and Wang, C. 2008. Chapter 55: Particle size distribution. In: Soil Sampling and Methods of Analysis. Carter, M.R. and Gregorich, E.G. (Eds). 2nd Edition. Pinawa, Manitoba; Boca Raton, FL: Canadian Society of Soil Science, CRC Press.

Lagacherie, P. 2008. Digital soil mapping: a state of the art. In: Hartemink, A.E., McBratney, A., Mendonça-Santos, M.L. (Eds). Digital Soil Mapping with Limited Data. Dordrecht; London: Springer.

Laurence, L., Heung, B., Strom, H., Stiles, K., Burton, D. 2023. Towards a cost-effective framework for estimating soil nitrogen pools using pedotransfer functions and machine learning. Geoderma, **440:** 116692.

Li, Y., Li, Z., Cui, S., Zhang, Q. 2020. Trade-off between soil pH, bulk density and other soil physical properties under global no-tillage agriculture. Geoderma, **361:** 114099.

Liaw, A., and Wiener, M. 2002. Classification and regression by randomForest. R News, **2/3:** 18-22.

Lombardo, L., Saia, S., Schillaci, C., Mai, P.M., Huser, R. 2018. Modeling soil organic carbon with quantile regression: Dissecting predictors' effects on carbon stocks. Geoderma, **318:** 148-159.

López López, P., Verkade, J.S., Weerts, A.H. and Solomatine, D.P. 2014. Alternative configurations of quantile regression for estimating predictive uncertainty in water level forecasts for the upper Severn River: a comparison. Hydrology and Earth System Sciences, **18:** 311-3428.

Malone, B.P., McBratney, A.B., Minasny, B. 2011. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. Geoderma, **160:** 614-626.

Manrique, L.A., Jones, C.A., 1991. Bulk density of soils in relation to soil physical and chemical properties. Soil Sci. Soc. Am. J., **55:** 476–481.

Martin, M.P., Lo Seen, D., Boulonne, L., Jolivet, C., Nair, K.M., Bourgeon, G., Arrouays, D. 2009. Optimizing pedotransfer functions for estimating soil bulk density using boosted regression trees. Soil Sci. Soc. Am. J., **73:** 485-493.

McBratney, A.B., Minasny, B., Cattle, S.R., and Vervoort, R.W. 2002. From pedotransfer functions to soil inference systems. Geoderma, **109:** 41-73.

Mehler, K., Schöning, I., Berli, M. 2014. The importance of rock fragment density for the calculation of soil bulk density and soil organic carbon stocks. Soil Sci. Soc. Am. J., **78:** 1186-1191.

Minasny, B., and McBratney, A.B. 2002. Uncertainty analysis for pedotransfer functions. European Journal of Soil Science, **53:** 417-429.

Minasny, B. and Hartemink, A.E. 2011. Predicting soil properties in the tropics. Earth-Science Reviews, **106:** 52-62.

Moghaddam, D.D., Rahmati, O., Panahi, M., Tiefenbacher, J., Darabi, H., Haghizadeh, A., Haghighi, A.T., Nalivan, O.A., Bui, D.T. 2020. The effect of sample size on different machine learning models for groundwater potential mapping in mountain bedrock aquifers. Catena, **187:** 104421.

Muthusamy, M., Godiksen, P.N., Madsen, H. 2016. Comparison of different configurations of quantile regression in estimating predictive hydrological uncertainty. Procedia Engineering, **154:** 513-520.

Nanko, K., Ugawa, S., Hashimoto, S., Imaya, A., Kob ayashi, M., Sakai, H., Ishizuka, S., Miura, S., Tanaka, N., Takahashi, M., Kaneko, S. 2014. A pedotransfer function for estimating bulk density of forest soil in Japan affected by volcanic ash. Geoderma, **213:** 36-45.

Nasta, P., Palladino, M., Sica, B., Pizzolante, A., Trifuoggi, M., Toscanesi, M., Giarra, A., D'Auria, J., Nicodemo, F., Mazzitelli, C., Lazzaro, U., Di Fiore, P., Romano, N. 2020. Evaluating pedotransfer functions for predicting soil bulk density using hierarchical mapping information in Campania, Italy. Geoderma Regional, **21:** e00267

Padarian, J., Minasny, B., and McBratney, A.B. 2020. Machine learning and soil sciences: a review aided by machine learning tools. Soil, **6:** 35-52.

Patton, N.R., Lohse, K.A., Seyfried, M., Will, R., Benner, S.G. 2019. Lithology and coarse fraction adjusted bulk density estimates for determining total organic carbon stocks in dryland soils. Geoderma, **337:** 844-852.

Pereira, O.J.R., Montes, C.R., Lucas, Y., Melfi, A.J., 2016. Evaluation of pedotransfer equations to predict deep soil carbon stock in tropical podzols compared to other soils of the Brazilian Amazon forest. In: Hartemink, A.E., Minasny, B. (Eds.), Digital Soil Morphometrics, Progress in Soil Science, 331–349. Springer International Publishing, Switzerland.

Qiao, J., Zhu, Y., Jia, X., Huang, L., Shao, M., 2019. Development of pedotransfer functions for predicting the bulk density in the critical zone on the Loess Plateau, China. J. Soil. Sediment., **19:** 366–372.

Qin, L., Lin, L., Ding, S., Yi, C., Chen, J., Tian, Z. Evaluation of pedotransfer functions for predicting particle density of soils with low organic matter contents. Geoderma, **416:** 115812.

Quesnel, H., and Suttie, K. 1983. Data entry procedures for laboratory forms (BCSIS Volume 3). Information Services Branch, British Columbia Ministry of Forests. Publication No. R28-82055.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Rahmati, O., Choubin, B., Fathabadi, A., Coulon, F., Soltani, E., Shahabi, H., Mollaefar, E., Tiefenbacher, J., Cipullo, S., Bin Ahmad, B., Bui, D.T. 2019. Predicting uncertainty of machine learning models for modelling nitrate pollution of groundwater using quantile regression and UNEEC methods. Science of the Total Environment, 688: 855-866.

Ramcharan, A., Hengl, T., Beaudette, D., and Wills, S. 2017. A soil bulk density pedotransfer function based on machine learning: a case study with the NCSS soil characterization database. Soil Sci. Soc. Am. J., **81:** 1279-1287.

Reidy, B., Simo, I., Sills, P., and Creamer, R.E. 2016. SOIL, **2:** 25-39.

Rossiter, D. 2008. Digital soil mapping as a component of data renewal for areas with sparse soil data infrastructures. In: Hartemink, A.E., McBratney, A., Mendonça-Santos, M.L. (Eds). Digital Soil Mapping with Limited Data. Dordrecht; London: Springer.

Schillaci, C., Perego, A., Valkama, E., Märker, M., Saia, S., Veronesi, F., Lipani, A., Lombardo, L., Tadiello, T., Gamper, H.A., Tedone, L., Moss, C., Pareja-Serrano, E., Amato, G., Kühl, K., Dămătîrcă, C., Cogato, A., Mzid, N., Eeswaran, R., Rabelo, M., Sperandio, G., Bosino, A., Bufalini, M., Tunçay, T., Ding, J., Fiorentini, M., Tiscornia, G., Conradt, S., Botta, M., Acutis, M. 2021. New pedotransfer approaches to predict soil bulk density using WoSIS soil data and environmental covariates in Mediterranean agro-ecosystems. Science of the Total Environment, **780:** 146609.

Schmidinger, J., and Heuvelink, G.B.M. 2023. Validation of uncertainty predictions in digital soil mapping. Geoderma, **437:** 116585.

Sequeira, C.H., Wills, S.A., Seybold, C.A., and West, L.T. 2014. Predicting soil bulk density for incomplete databases. Geoderma, **213:** 64-73.

Sevastas, S., Gasparatos, D., Botsis, D., Siarkos, I., Diamantaras, K.I. and Bilas, G. 2018. Predicting bulk density using pedotransfer functions for soils in the Upper Anthemountas basin, Greece. Geoderma Regional, **14:** e00169.

Silatsa, F.B.T., Yemefack, M., Tabi, F.O., Heuvelink, G.B.M., Leenaars, J.G.B. 2020. Assessing countrywide soil organic carbon stock using hybrid machine learning modelling and legacy soil data in Cameroon. Geoderma, **367:** 114260.

Soil Classification Working Group, 1998. The Canadian System of Soil Classification. NRC Research Press.

Somarathna, P.D.S.N., Minasny, B., Malone, B.P. 2017. More data or a better model? Figuring out what matters most for the spatial prediction of soil carbon. Soil Sci. Soc. Am. J., **81:** 1413-1426.

Song, G., Li, L., Pan, G., Zhang, Q., 2005. Topsoil organic carbon storage of China and its loss by cultivation. Biogeochemistry, **74:** 47–62.

Staffa, S.J., Kohane, D.S., Zurakowski, D. 2019. Quantile regression and its applications: a primer for anesthesiologists. Anesthesia Analgesia, **128:** 820-830.

Szabó, B., Weynants, M., Weber, T.K.D. 2021. Updated European hydraulic pedotransfer functions with communicated uncertainties in the predicted variables (euptfv2). Geosci. Model Dev., **14:** 151-175.

Tamminen, P., Starr, M., 1994. Bulk density of forested mineral soils. Silva Fennica, **28(1):** 53–60.

Tranter, G., Minasny, B., McBratney, A.B. 2010. Estimating pedotransfer function prediction limits using fuzzy k-means with extragrades. Soil Sci. Soc. Am. J., **74:** 1967-1975.

Tomasella, J., Hodnett, M.G., 1998. Estimating soil water retention characteristics from limited data in Brazilian Amazonia. Soil Sci., **163(3):** 190–202.

Van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., Montzka, C., Nemes, A., Pachepsky, Y.A., Padarian, J., Schaap, M.G., Tóth, B., Verhoef, A., Vanderborght, J., van der Ploeg, M.J., Weihermüller, L., Zacharias, S., Zhang, Y., and Vereecken, H. 2017. Pedotransfer functions in Earth System Science: Challenges and perspectives. Reviews of Geophysics, **55:** 1199-1256.

Wang, Y., Shao, M., Liu, Z., and Zhang, C. 2014. Prediction of bulk density of soils in the Loess Plateau region of China. Surv. Geophys. **35:** 395-413.

Wu, T., Wu, Q., Zhuang, Q., Li, Y., Yao, Y., Zhang, L., and Xing, S. 2022. Optimal sample size for SOC content prediction for mapping using the random forest in cropland in northern Jiangsu, China. Eurasian Soil Science, **55(12):** 1689-1699.

Yi., X.S., Li, G.S., Yin, Y.Y. 2016. Pedotransfer functions for estimating soil bulk density: A case study in the Three-River Headwater region of Qinghai Province, China. Pedosphere, **26(3):** 362-373.

Zhang, J., Schmidt, M.G., Heung, B., Bulmer, C.E., Knudby, A. 2022. Using an ensemble learning approach in digital soil mapping of soil pH for the Thompson-Okanagan region of British Columbia. Can. J. Soil Sci., **102:** 579-596.

Zihao, H., Shaofei, J., and Ku, W. 2022. Application of machine learning methods for estimation soil bulk density. 2022 2nd Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS), Shenyang, China, pp. 194-198.

3.7. Tables

Table 3.1. Summary statistics of Dataset (n = 101,722)

Attribute	Min	1st Quarter	Median	Mean	3rd Quarter	Max
BD	0.30	1.11	1.35	1.41	1.69	2.62
Depth	0.00	11.00	34.50	40.86	65.00	1,322.50
OC	0.00	0.71	1.53	2.58	3.25	16.96
CEC	0.00	7.20	12.90	15.59	20.80	160.90
TN	0.00	0.04	0.07	0.14	0.13	57.00
pH	0.84	5.30	5.80	6.03	6.70	11.60
CF	0.00	0.00	10.00	22.62	45.00	100.00
Sand	0.00	31.00	46.72	46.38	62.50	99.00
Silt	0.00	28.00	38.50	39.02	50.00	98.00
Clay	0.00	6.53	11.50	14.65	18.50	96.21
Order	Brunisolic (36,264), Chernozemic (2,834), Gleysolic (12,967), Luvisolic (6,744), Podzolic (29,384), Regosolic (3,377), Solonetzic (41).					
PM	Bedrock (81), Colluvial (6,560), Eolian (1,626), Fluvial (17,969), Glacial Fluvial (10,184), Glacial Lacustrine (1,463), Lacustrine (1,578), Marine (29,378), Morainal (25,187), Organic (925), Saprolite (93), Volcanic (124)					
Textural Class	Clay (579), Clay Loam (2,425), Heavy Clay (121), Loam (12,937), Loamy Sand (12,591), Sand (9,801), Sandy Clay (73), Sandy Clay Loam (1,954), Sandy Loam (26,117), Silt (370), Silty Clay (1,266), Silty Clay Loam (5,245), Silty Loam (12,376)					

Table 3.2. Number and percentage of horizons with measured values (of n = 101,722)

Attribute	Number of Observations	Percent Horizons with Observations
BD	1,450	1.43
Depth	101,722	100.00
OC	17,336	17.04
CEC	14,158	13.92
TN	14,877	14.63
pH	22,322	21.94
CF	100,113	98.42
Sand	9,951	9.78
Silt	9,887	9.72
Clay	9,916	9.75
Order	91,611	90.06
PM	95,169	93.56
Textural Class	85,856	84.40

Table 3.3. Top 40 best performing models

Rank	Model	CCC	RMSE
1	BD = f(depth + CF + OC + pH + sand + clay)	0.92	0.10
2	BD = f(depth + CF + OC + sand + clay)	0.92	0.11
3	BD = f(depth + CEC + CF + pH + sand + clay)	0.91	0.11
4	BD = f(depth + CF + pH + sand + clay)	0.91	0.12
5	BD = f(depth + OC + order + sand + clay)	0.91	0.11
6	BD = f(depth + OC + order + pH + sand + clay)	0.91	0.11
7	BD = f(depth + CEC + OC + sand + clay)	0.91	0.11
8	BD = f(depth + CF + OC + order + pH + sand + clay)	0.91	0.10
9	BD = f(depth + OC + pH + sand + clay)	0.91	0.11
10	BD = f(depth + OC + sand + clay)	0.91	0.12
11	BD = f(depth + CF + OC + order + sand + clay)	0.91	0.11
12	BD = f(depth + CEC + pH + sand + clay)	0.91	0.12
13	BD = f(depth + CEC + OC + pH)	0.91	0.14
14	BD = f(depth + CEC + OC + order + sand + clay)	0.91	0.11
15	BD = f(depth + CEC + CF + OC + pH + sand + clay)	0.91	0.11
16	BD = f(depth + CF + pH + sand + clay + TN)	0.91	0.11
17	BD = f(depth + CEC + OC + pH + sand + clay)	0.91	0.11
18	BD = f(depth + CEC + CF + OC + sand + clay)	0.91	0.11
19	BD = f(depth + CEC + CF + sand + clay + TN)	0.90	0.11
20	BD = f(depth + CF + order + pH + sand + clay)	0.90	0.12
21	BD = f(depth + CEC + CF + OC + pH + TN)	0.90	0.14

Rank	Model	CCC	RMSE
22	BD = f(depth + CF + sand + clay + TN)	0.90	0.11
23	BD = f(depth + CEC + OC + pH + TN)	0.90	0.14
24	BD = f(depth + CEC + CF + pH + sand + clay + TN)	0.90	0.11
25	BD = f(depth + CEC + CF + OC + sand + clay + TN)	0.90	0.11
26	BD = f(depth + CEC + CF + OC + pH)	0.90	0.15
27	BD = f(depth + OC + sand + clay + TN)	0.90	0.12
28	BD = f(depth + CEC + CF + order + pH + sand + clay)	0.90	0.12
29	BD = f(depth + CF + OC + pH + sand + clay + TN)	0.90	0.11
30	BD = f(depth + CEC + pH)	0.90	0.15
31	BD = f(depth + CF + OC + pH)	0.90	0.15
32	BD = f(depth + CF + OC + pH + TN)	0.90	0.15
33	BD = f(depth + CF + OC + sand + clay + TN)	0.90	0.12
34	BD = f(depth + CEC + OC + sand + clay + TN)	0.90	0.12
35	BD = f(depth + OC + pH)	0.90	0.15
36	BD = f(depth + CEC + OC)	0.90	0.15
37	BD = f(depth + CEC + OC + TN)	0.90	0.15
38	BD = f(depth + OC + pH + sand + clay + TN)	0.90	0.11
39	BD = f(depth + CEC + CF + OC + TN)	0.90	0.15
40	BD = f(depth + CEC + pH + TN)	0.90	0.15

Table 3.4. 10 worst performing models

Rank	Model	CCC	RMSE
503	BD = f(depth + PM + textural class)	0.59	0.31
504	BD = f(depth + order + textural class)	0.58	0.29
505	BD = f(depth + CEC + TN + order + textural class)	0.58	0.25
506	BD = f(depth + order + PM)	0.58	0.29
507	BD = f(depth + OC + order + textural class)	0.56	0.25
508	BD = f(depth + TN + order + PM)	0.54	0.26
509	BD = f(depth + order)	0.53	0.31
510	BD = f(depth + textural class)	0.53	0.33
511	BD = f(depth + PM)	0.52	0.32
512	BD = f(depth + TN + order + textural class)	0.51	0.26

Table 3.5. Models used for gap filling bulk density values

Order Used	Rank	Model	CCC	No. of training horizons	Horizons Filled
1	1	BD = f(depth + CF + OC + pH + sand + clay)	0.92	381	6,379
2	2	BD = f(depth + CF + OC + sand + clay)	0.92	385	709
3	3	BD = f(depth + CEC + CF + pH + sand + clay)	0.91	369	460
4	4	BD = f(depth + CF + pH + sand + clay)	0.91	454	964
5	9	BD = f(depth + OC + pH + sand + clay)	0.91	415	190
6	12	BD = f(depth + CEC + pH + sand + clay)	0.91	396	37
7	13	BD = f(depth + CEC + OC + pH)	0.91	871	7,347
8	30	BD = f(depth + CEC + pH)	0.90	971	2,379
9	35	BD = f(depth + OC + pH)	0.90	986	169
10	53	BD = f(depth + CEC + CF + OC + order + pH + sand + clay)	0.89	284	15
11	54	BD = f(depth + pH + sand + clay)	0.89	492	574
12	61	BD = f(depth + CEC + sand + clay)	0.89	402	2
13	74	BD = f(depth + CF + OC + TN)	0.89	872	21
14	89	BD = f(depth + CF + pH + TN)	0.89	876	245
15	122	BD = f(depth + sand + clay)	0.88	513	43
16	129	BD = f(depth + OC)	0.88	992	44
17	133	BD = f(depth + CEC)	0.88	979	17
18	167	BD = f(depth + pH + TN)	0.87	918	1
19	183	BD = f(depth + CF + order + pH + PM)	0.86	409	1,728
20	207	BD = f(depth + CF + TN)	0.86	879	10
21	244	BD = f(depth + CF + pH + PM)	0.85	1,095	883
22	265	BD = f(depth + CF + pH)	0.84	1,287	249
23	283	BD = f(depth + pH)	0.84	1,343	21
24	343	BD = f(depth + TN)	0.81	932	4
25	392	BD = f(depth + CF + order + PM + textural class)	0.78	561	64,069
26	417	BD = f(depth + CF)	0.75	1,392	13,187
27	496	BD = f(depth)	0.63	1,450	528

3.8. Figures

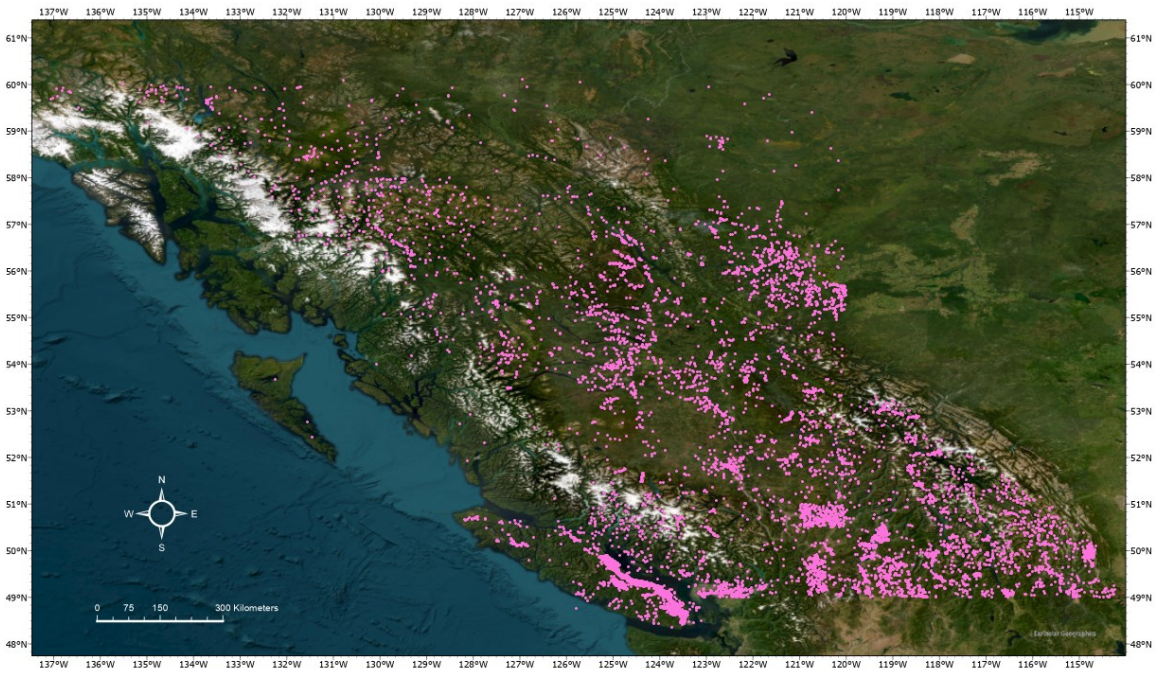


Figure 3.1. Sampling locations in the province of BC, Canada. Horizons from these sites were either used to train the models, or were filled with model predictions.

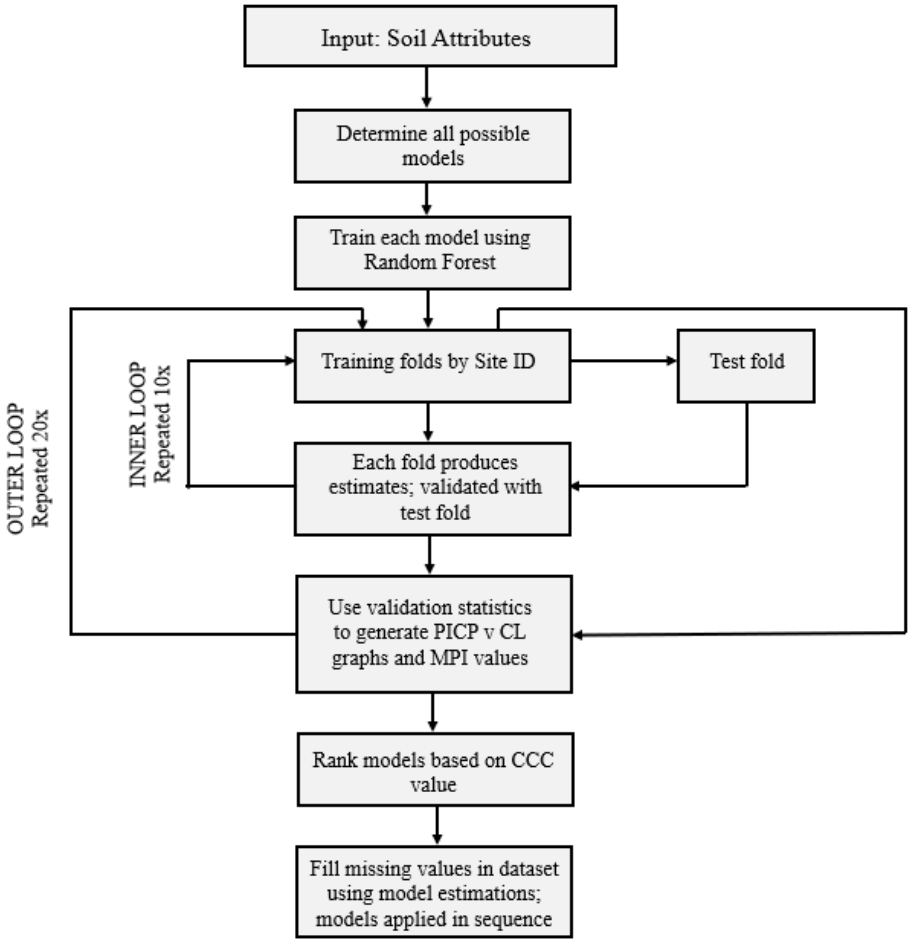


Figure 3.2. Schematic of methods

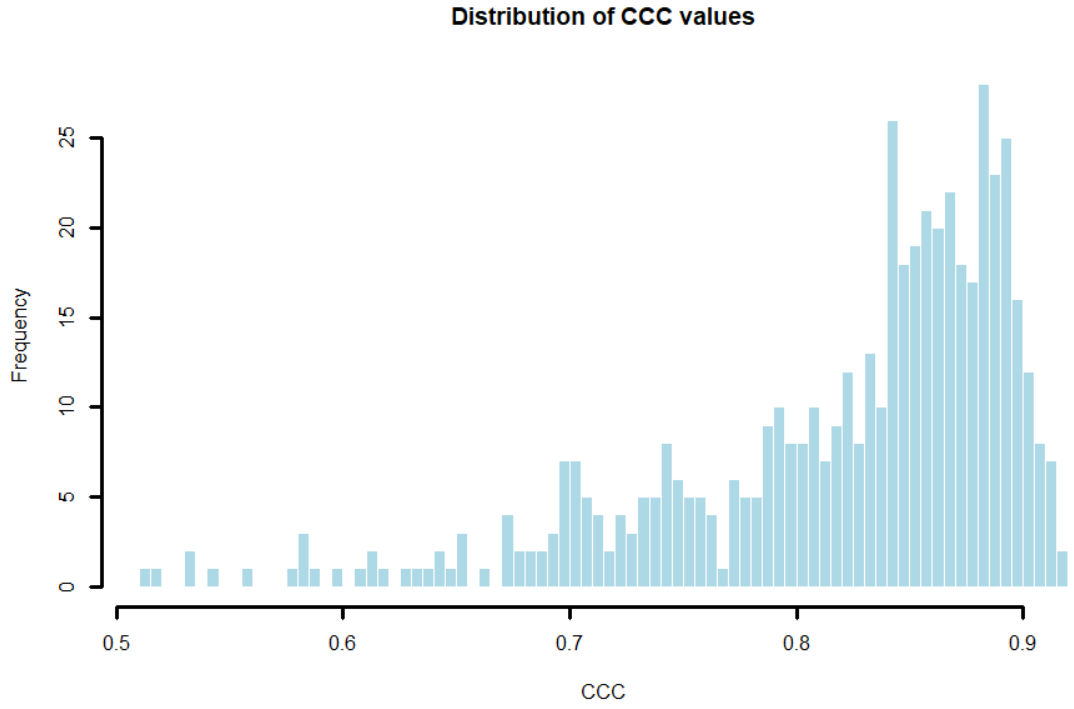


Figure 3.3. Distribution of CCC values after testing all 512 models.

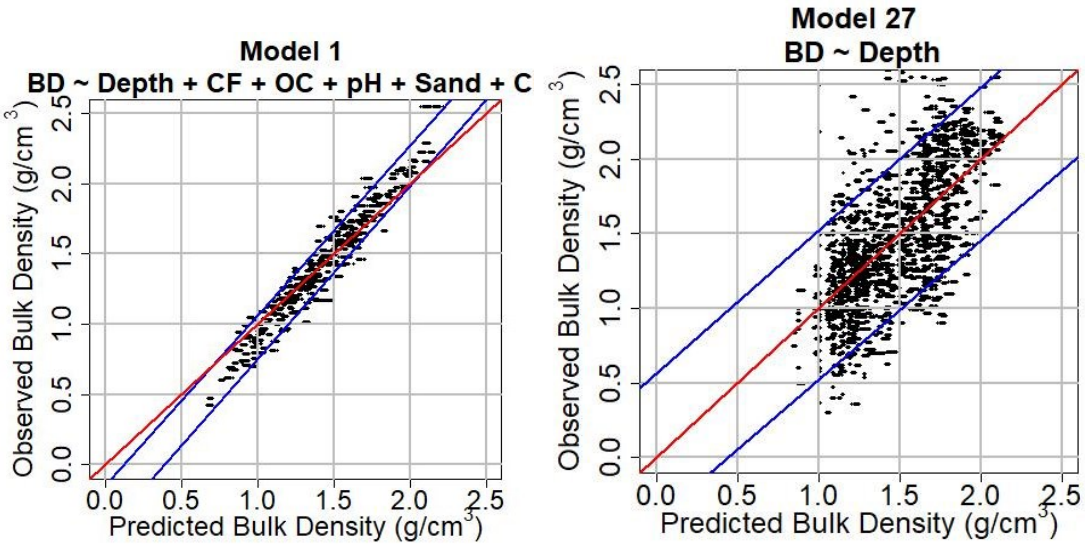


Figure 3.4. Observed vs Predicted plots for the first model applied (left), *BD = f(depth + CF + OC + pH + sand + clay)*; and last model applied (right), *BD = f(depth)*.

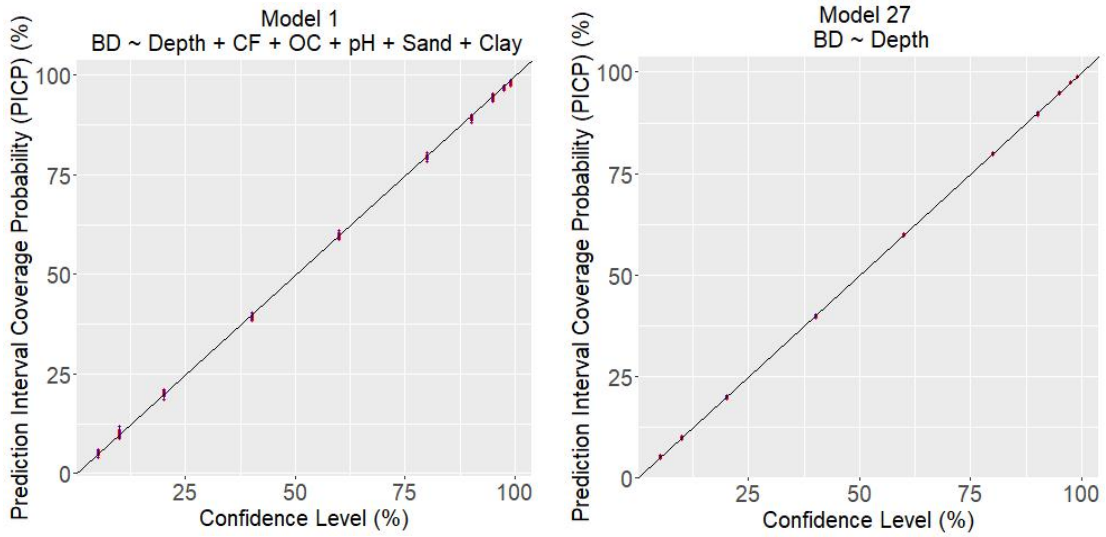


Figure 3.5. PICP vs CL graphs for the first model applied (left) and last model applied (right).

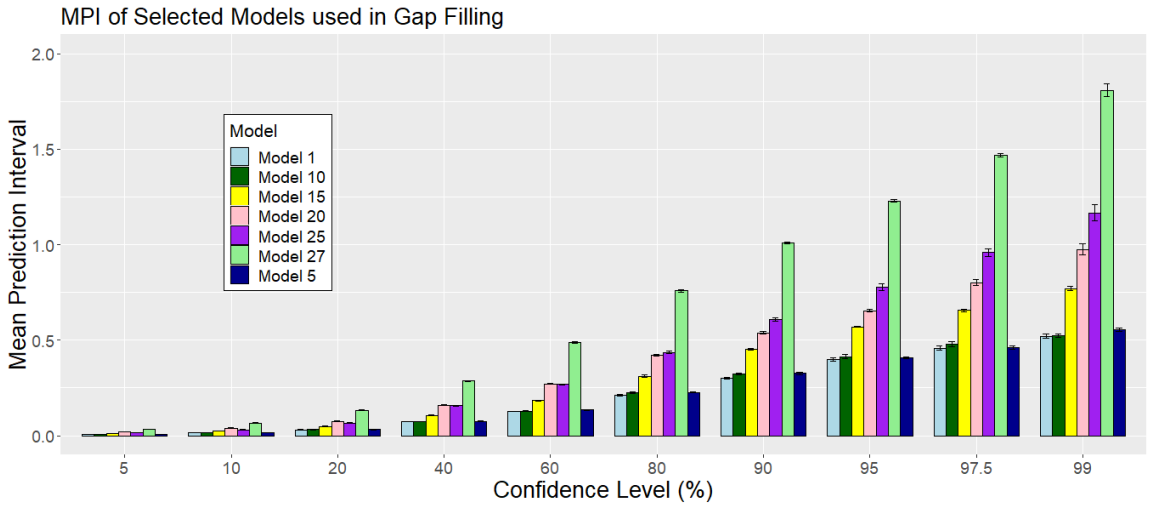


Figure 3.6. MPI values of selected models used in gap filling by confidence level.

3.9. Supplementary Figures

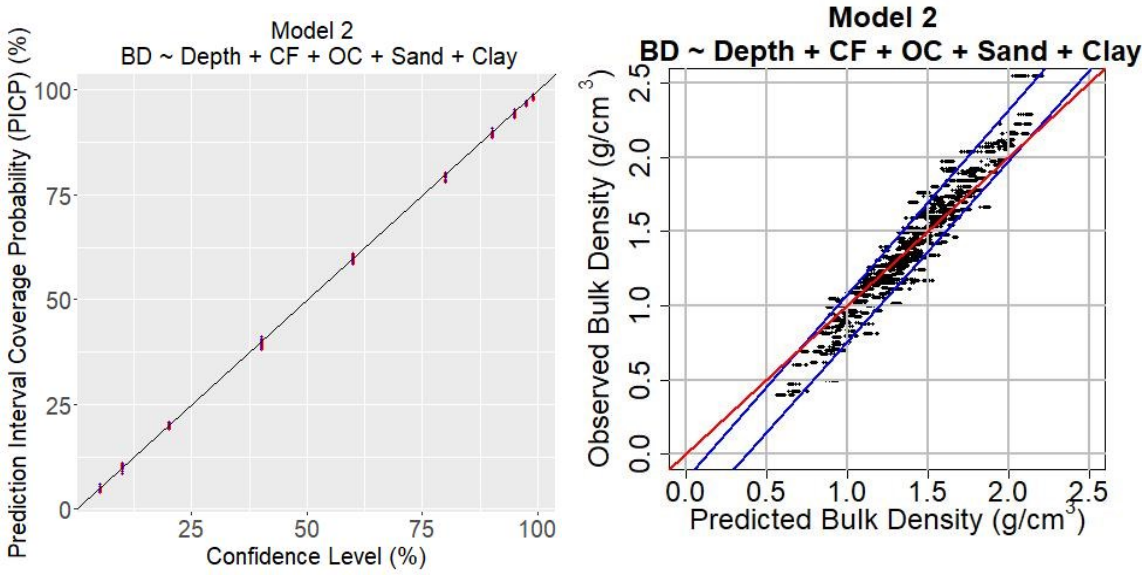


Figure 3.7. **Model 2, *BD = f(depth + CF + OC + sand + clay).* PICP vs CL graph, left; Observed vs predicted graph, right.**

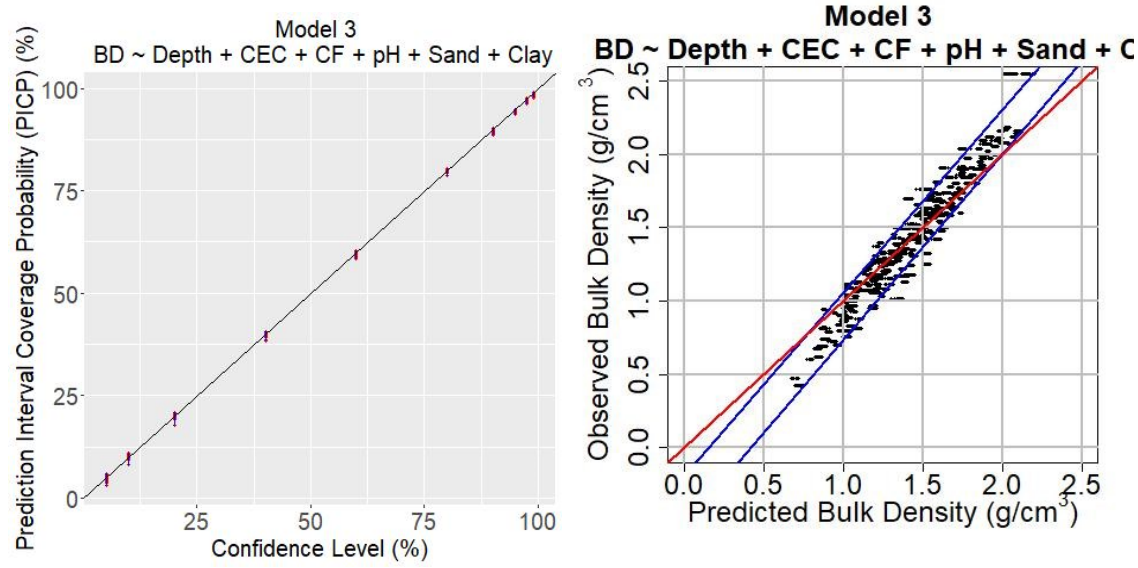


Figure 3.8. **Model 3, *BD = f(depth + CEC + CF + pH + sand + clay).* PICP vs CL graph, left; Observed vs predicted graph, right.**

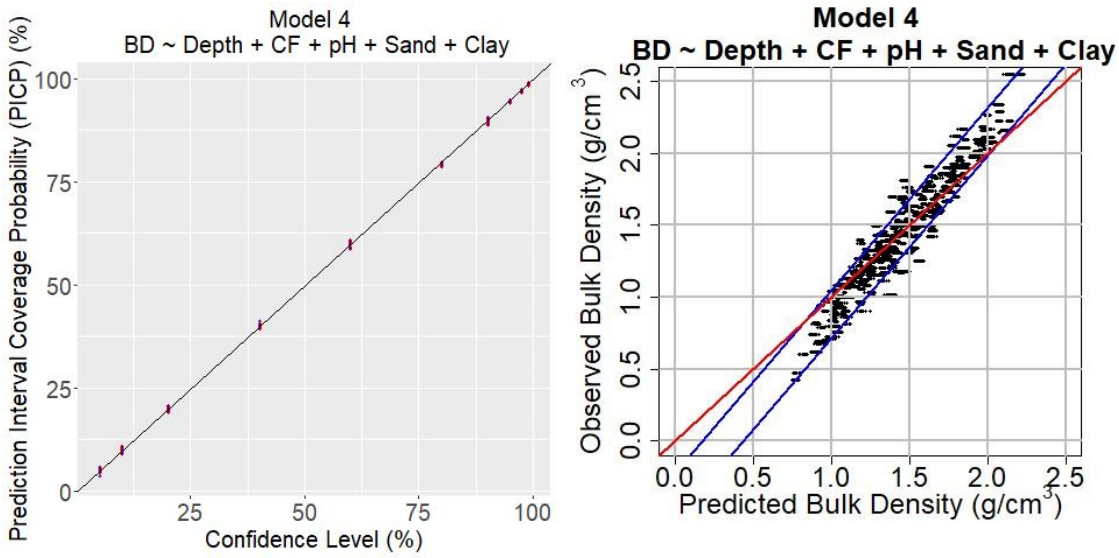


Figure 3.9. Model 4, *BD = f(depth + CF + pH + sand + clay).* PICP vs CL graph, left; Observed vs predicted graph, right.

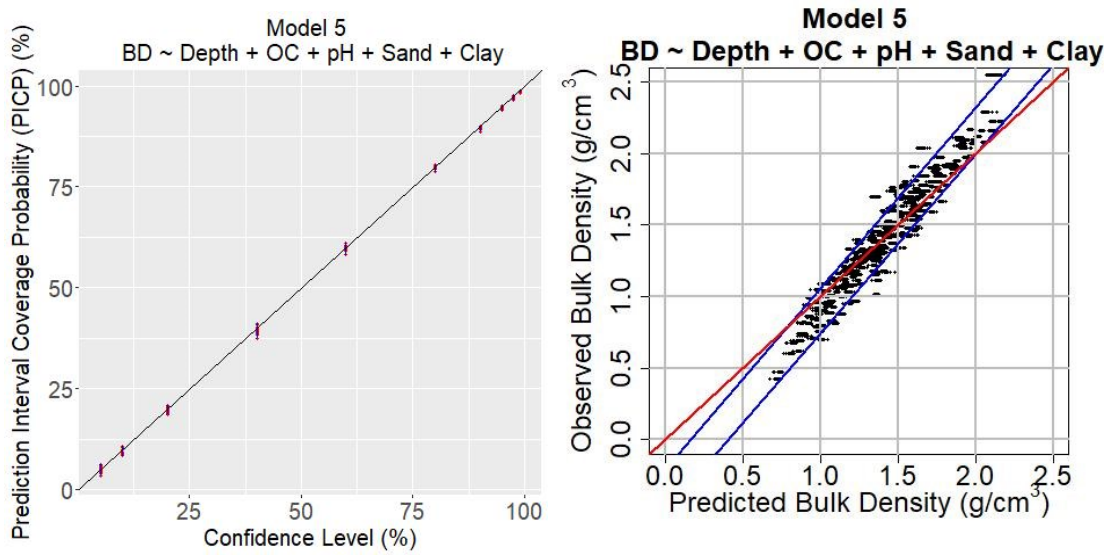


Figure 3.10. Model 5, *BD = f(depth + OC + pH + sand + clay).* PICP vs CL graph, left; Observed vs predicted graph, right.

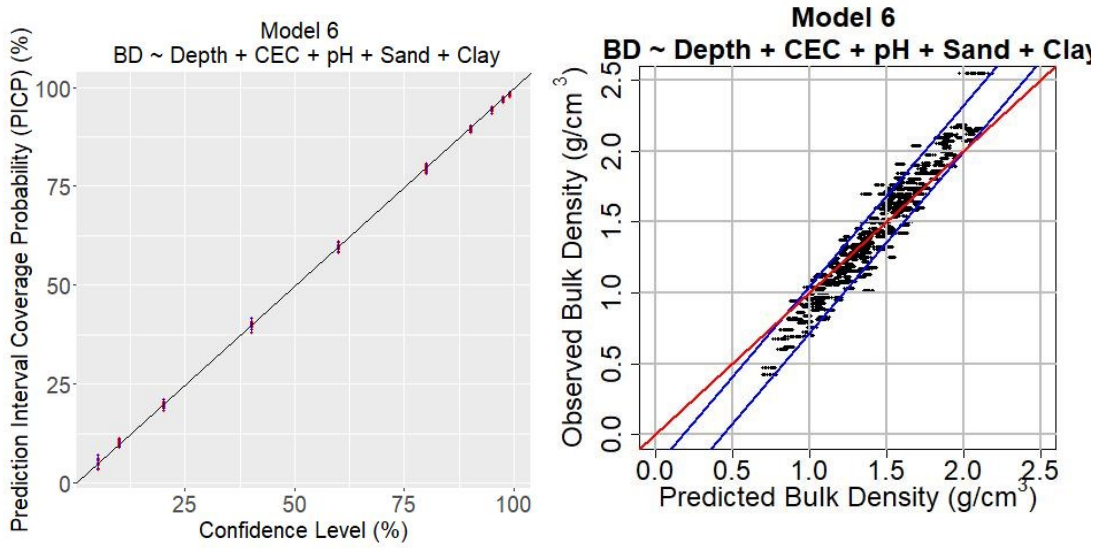


Figure 3.11. Model 6, *BD = f(depth + CEC + pH + sand + clay)*. PICP vs CL graph, left; Observed vs predicted graph, right.

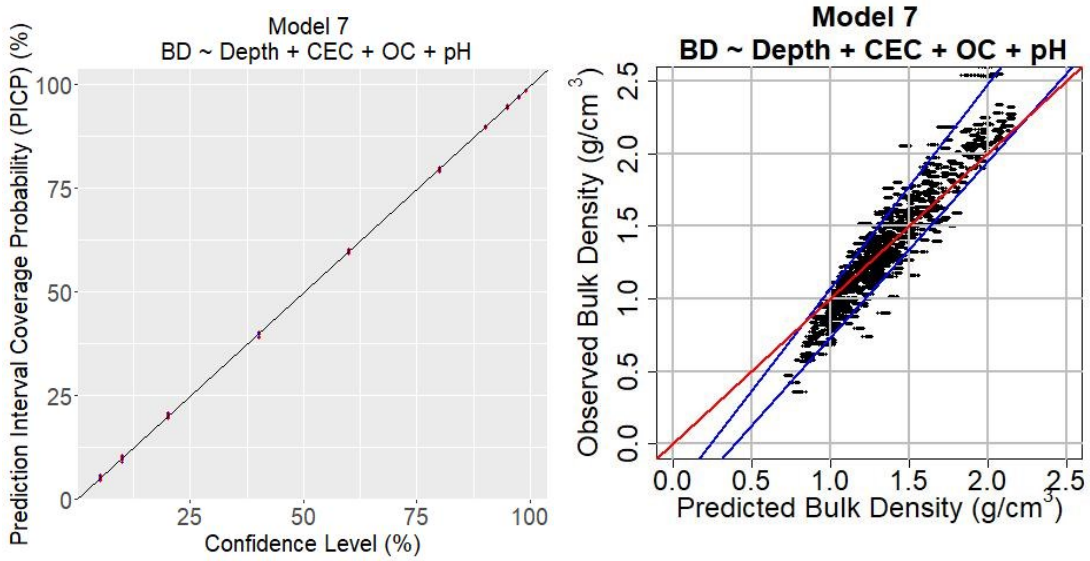


Figure 3.12. Model 7, *BD = f(depth + CEC + OC + pH)*. PICP vs CL graph, left; Observed vs predicted graph, right.

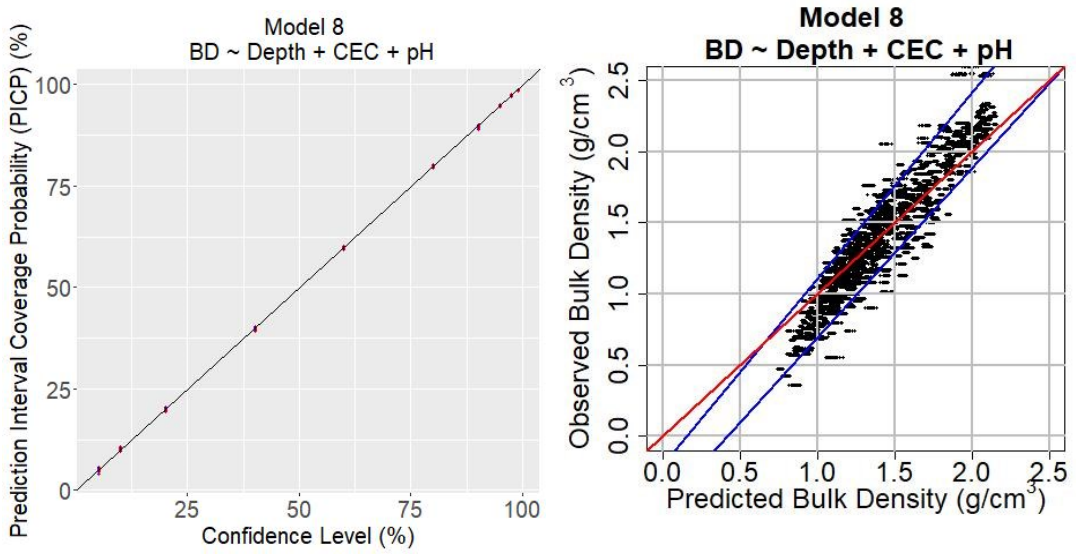


Figure 3.13. Model 8, *BD = f(depth + CEC + pH).* PICP vs CL graph, left; Observed vs predicted graph, right.

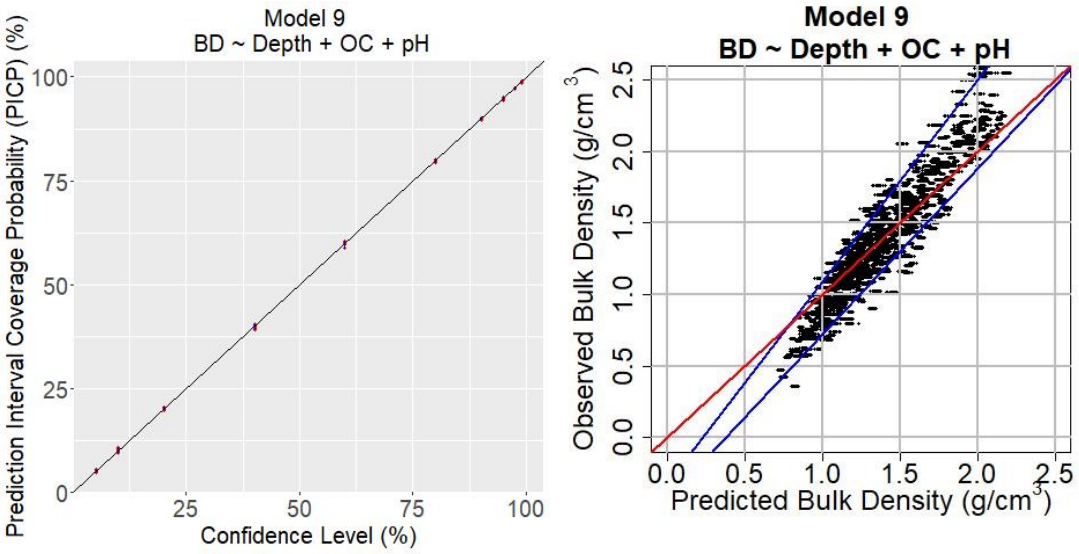


Figure 3.14. Model 9, *BD = f(depth + OC + pH).* PICP vs CL graph, left; Observed vs predicted graph, right.

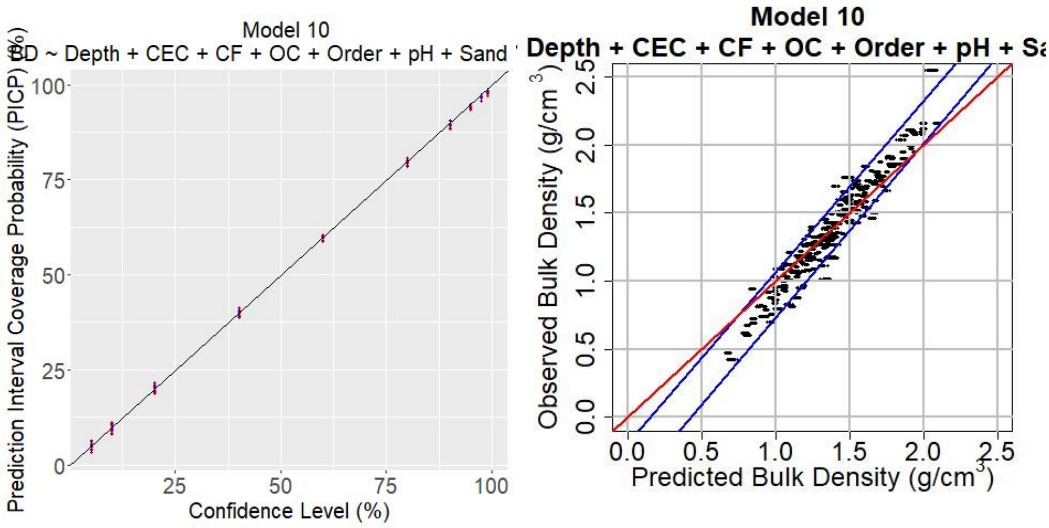


Figure 3.15. Model 10, *BD = f(depth + CEC + CF + OC + order + pH + sand + clay).* PICP vs CL graph, left; Observed vs predicted graph, right.

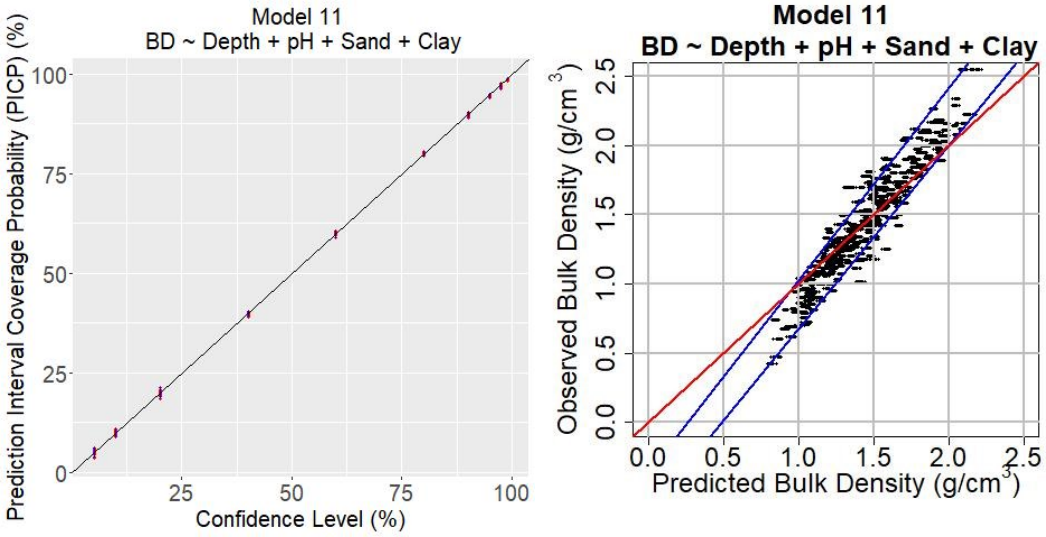


Figure 3.16. Model 11, *BD = f(depth + pH + sand + clay).* PICP vs CL graph, left; Observed vs predicted graph, right.

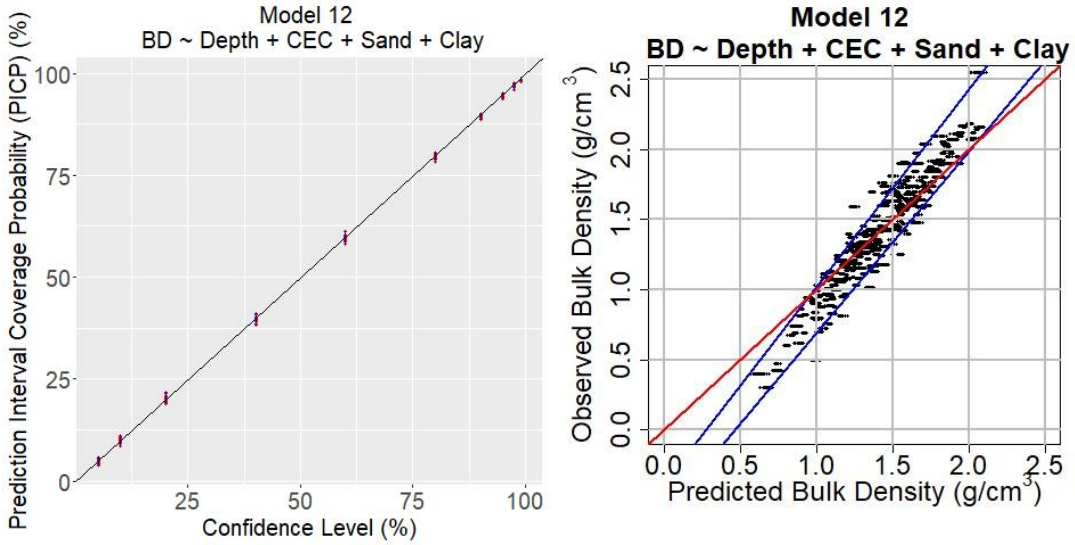


Figure 3.17. Model 12, *BD = f(depth + CEC + sand + clay).* PICP vs CL graph, left; Observed vs predicted graph, right.

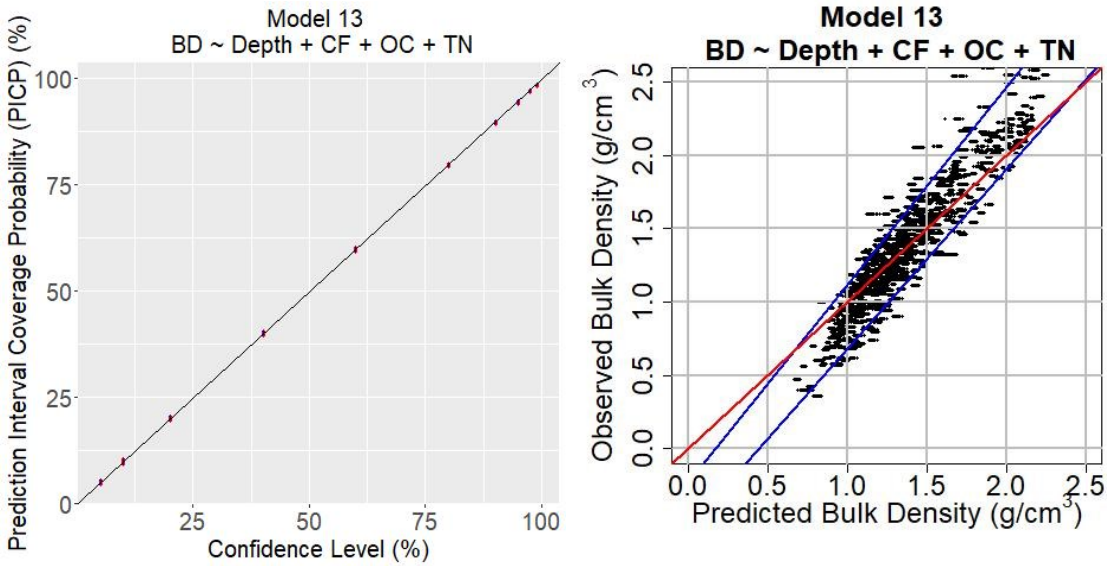


Figure 3.18. Model 13, *BD = f(depth + CF + OC + TN).* PICP vs CL graph, left; Observed vs predicted graph, right.

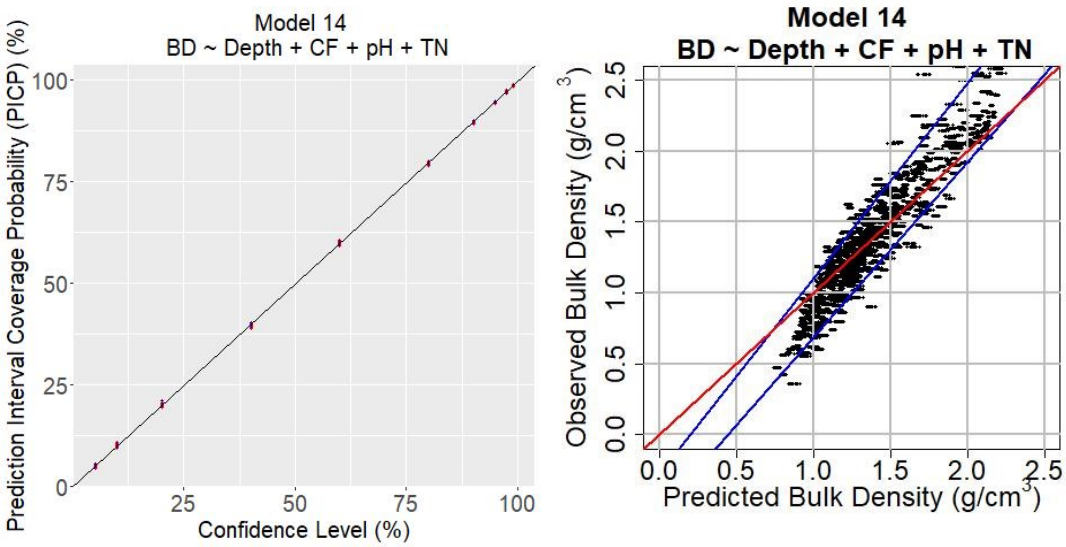


Figure 3.19. Model 14, *BD = f(depth + CF + pH + TN).* PICP vs CL graph, left; Observed vs predicted graph, right.

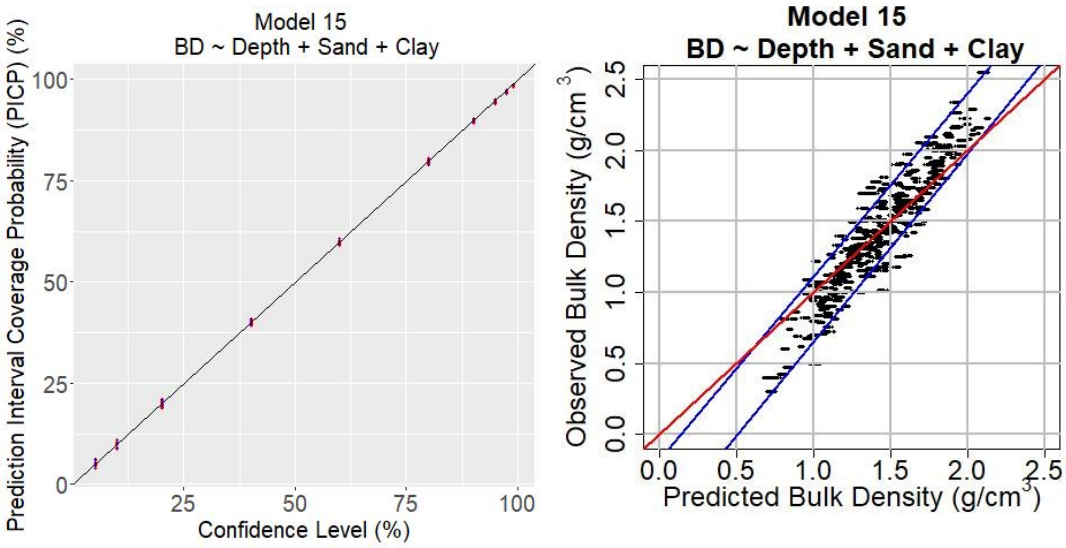


Figure 3.20. Model 15, *BD = f(depth + sand + clay).* PICP vs CL graph, left; Observed vs predicted graph, right.

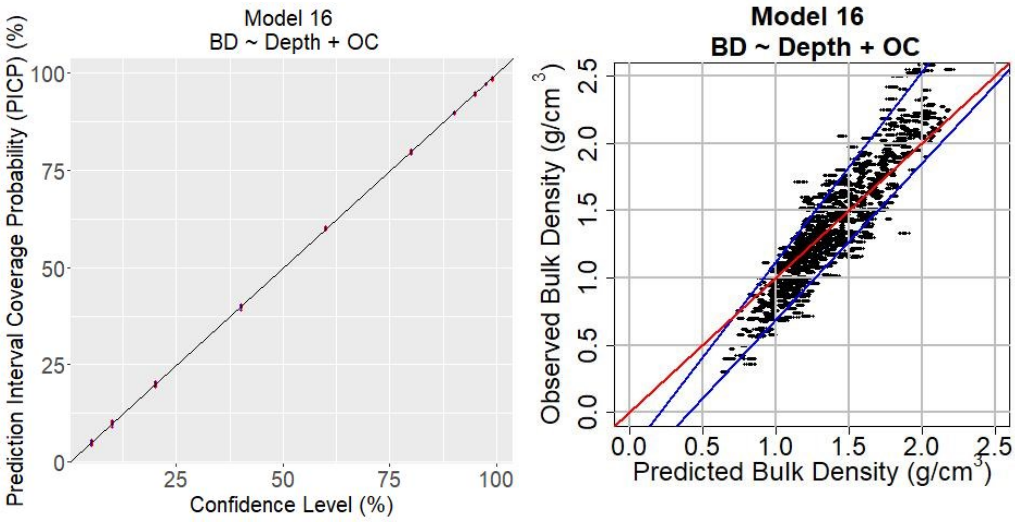


Figure 3.21. Model 16, *BD = f(depth + OC).* PICP vs CL graph, left; Observed vs predicted graph, right.

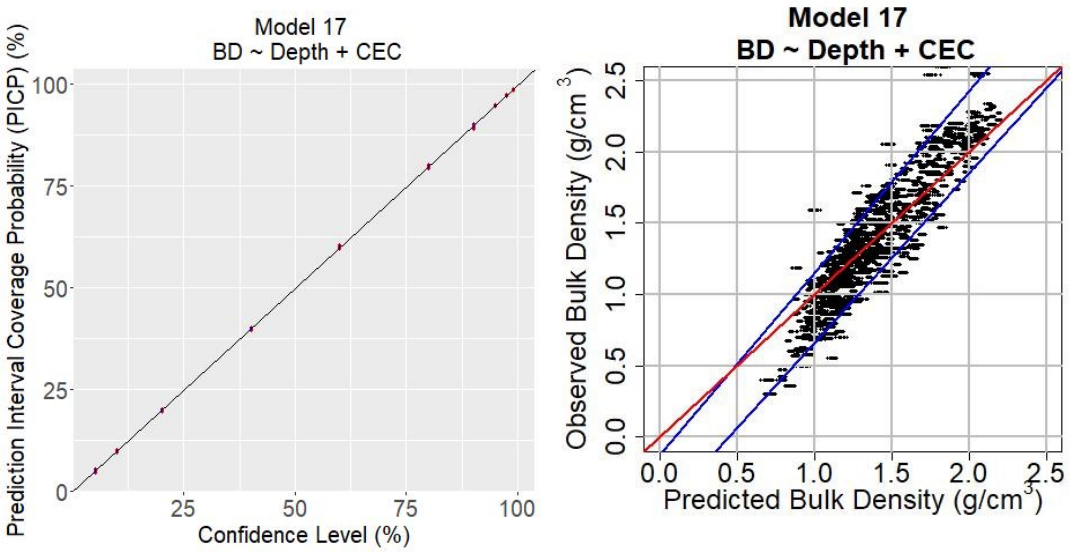


Figure 3.22. Model 17, *BD = f(depth + CEC).* PICP vs CL graph, left; Observed vs predicted graph, right.

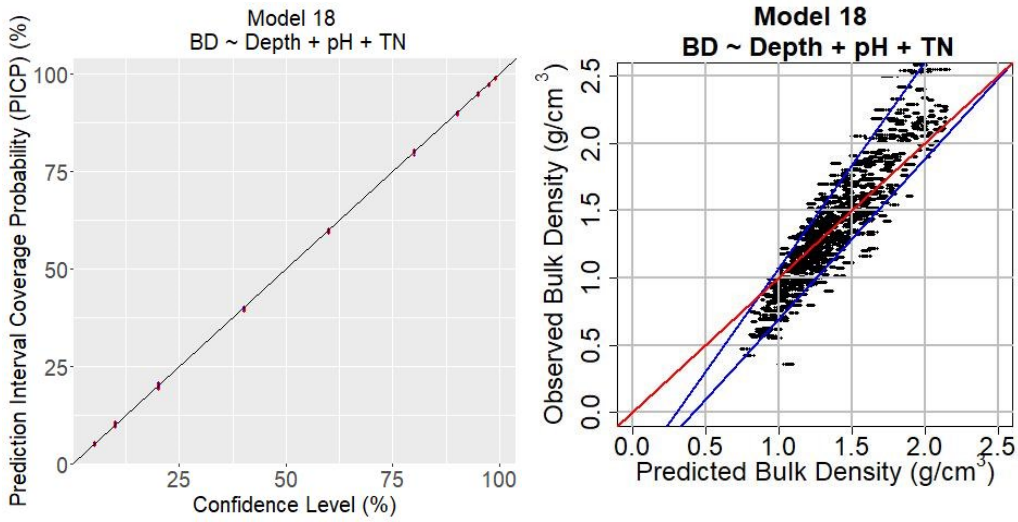


Figure 3.23. Model 18, *BD = f(depth + pH + TN).* PICP vs CL graph, left; Observed vs predicted graph, right.

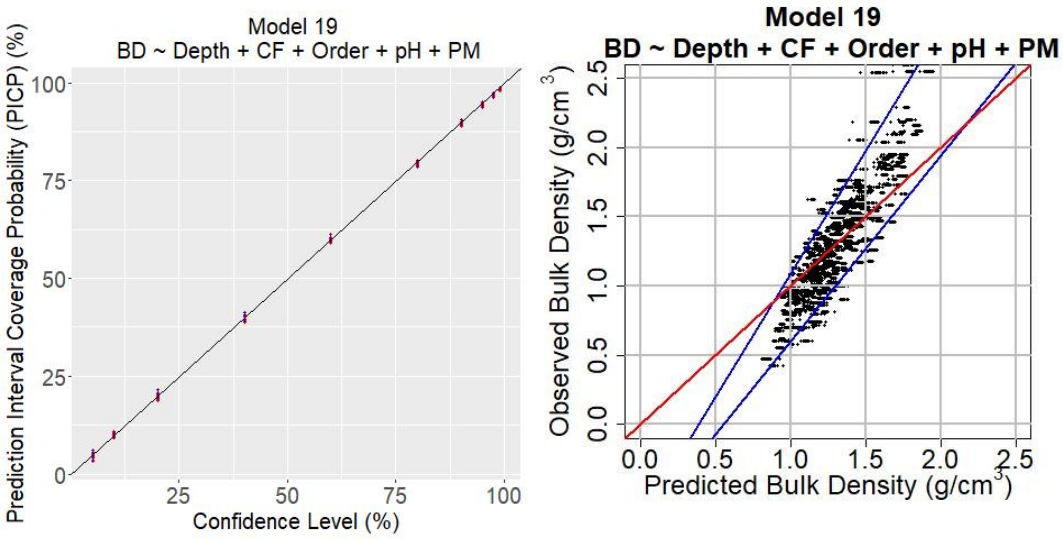


Figure 3.24. Model 19, *BD = f(depth + CF + order + pH + PM).* PICP vs CL graph, left; Observed vs predicted graph, right.

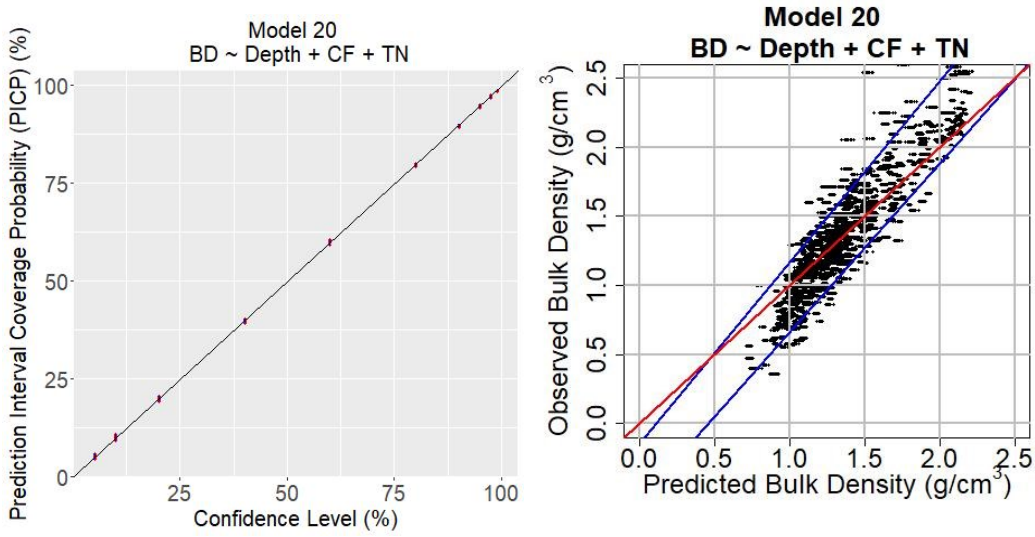


Figure 3.25. Model 20, *BD = f(depth + CF + TN).* PICP vs CL graph, left; Observed vs predicted graph, right.

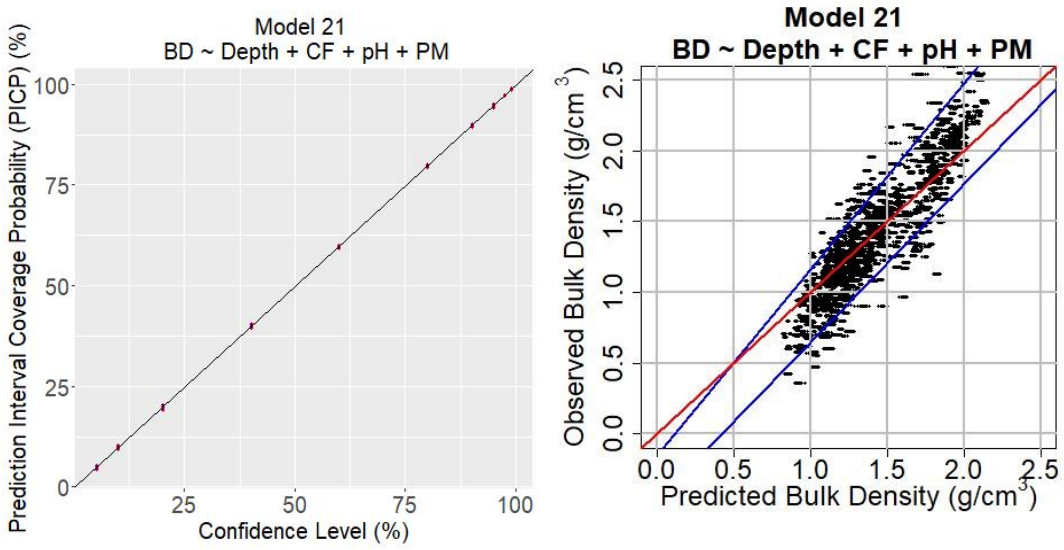


Figure 3.26. Model 21, *BD = f(depth + CF + pH + PM).* PICP vs CL graph, left; Observed vs predicted graph, right.

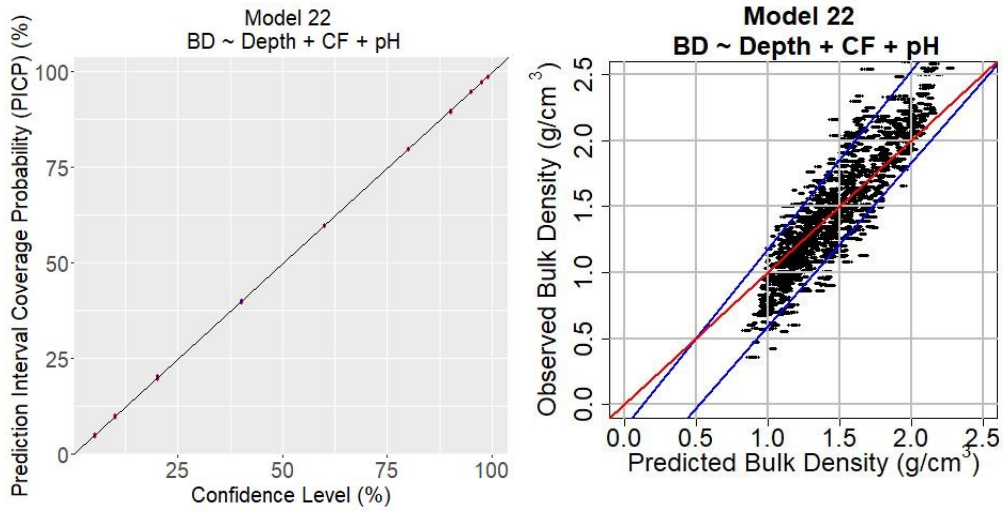


Figure 3.27. Model 22, *BD = f(depth + CF + pH).* PICP vs CL graph, left; Observed vs predicted graph, right.

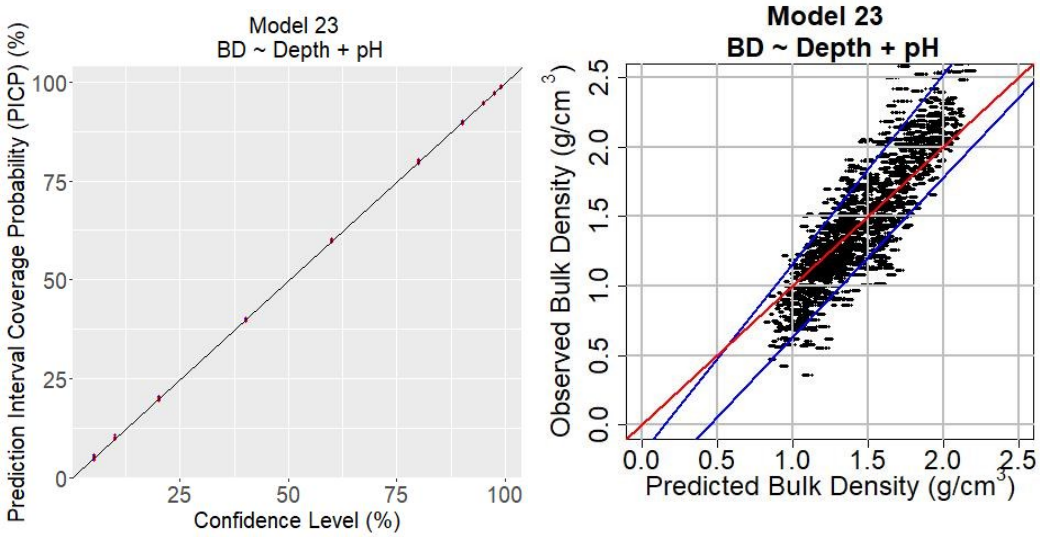


Figure 3.28. Model 23, *BD = f(depth + pH).* PICP vs CL graph, left; Observed vs predicted graph, right.

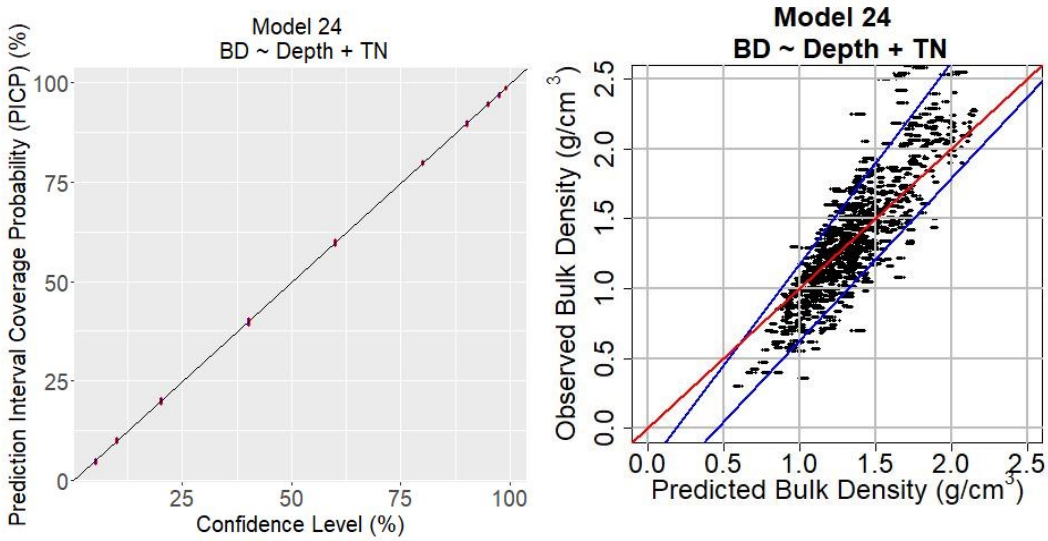


Figure 3.29. Model 24, *BD = f(depth + TN).* PICP vs CL graph, left; Observed vs predicted graph, right.

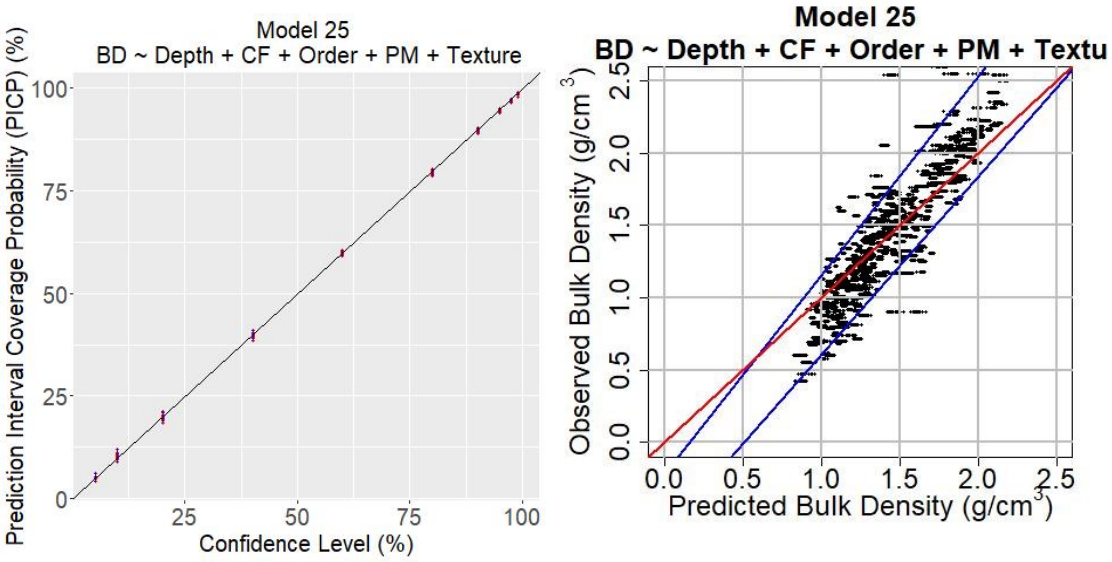


Figure 3.30. Model 25, *BD = f(depth + CF + order + PM + textural class).* PICP vs CL graph, left; Observed vs predicted graph, right.

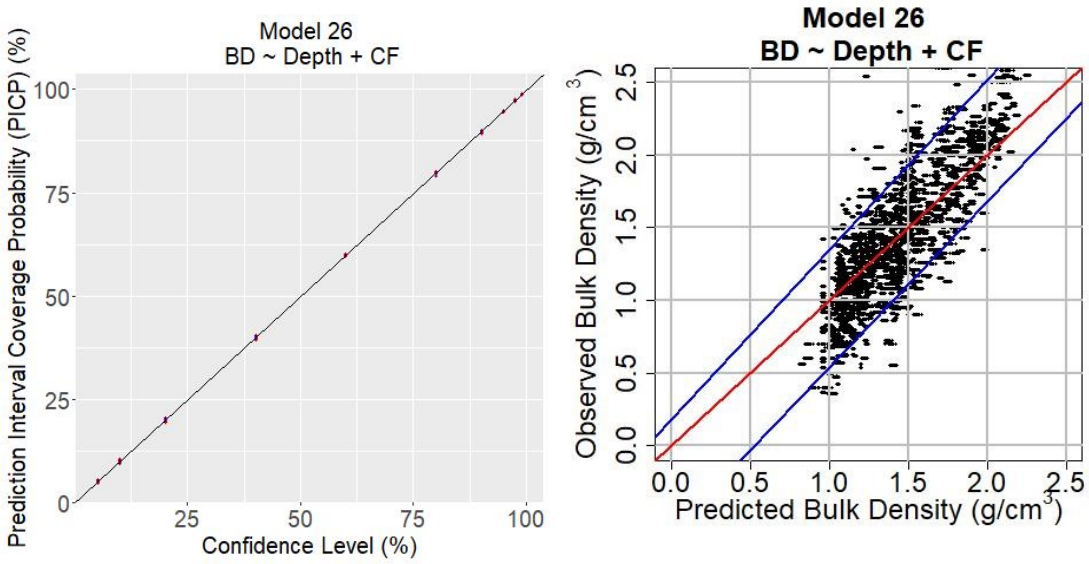


Figure 3.31. Model 26, *BD = f(depth + CF).* PICP vs CL graph, left; Observed vs predicted graph, right.

Chapter 4. Conclusion

4.1. Research Conclusions

Chapter 2 examined the use of nonlinear least squares to recalibrate existing PTFs on three regional datasets, using quantile regression to produce uncertainty estimates. The first objective of this thesis, identifying existing equation-based PTFs from the literature, was met when 73 PTFs which estimate bulk density were found. While there have been other studies which compared the performance of existing BD PTFs (Abdelbaki, 2018; Boschi et al., 2018; Casanova et al., 2016; De Vos et al., 2005; Han et al., 2012; Kaur et al., 2002; Nanko et al., 2014; Nasta et al., 2020; Reidy et al., 2016; Sevastas et al., 2018; Taulya et al., 2005; Vasiliniuc and Patriche, 2015; Yi et al., 2016) the focus of these studies was how existing PTFs performed on a regional dataset or specific type of soil; comparisons to a newly developed PTF were also made.

This study assessed the accuracy and uncertainty of existing PTFs on three regional datasets; the PTFs were then recalibrated on each dataset. A few other studies had recalibrated existing PTFs using an NLS approach (Chen et al., 2018; De Vos et al., 2005; Khodaverdiloo et al., 2022; Nanko et al., 2014) although they did not recalibrate as many PTFs as this study. Chen et al. (2018) recalibrated the six model forms defined by Nanko et al. (2014), which were all based on OC as the sole input variable; the RMSPE values of the results ranged from 0.186 to 0.278 g/cm^3. Nanko et al. (2014) reported a range of RMSPE values between 0.137 to 0.145 g/cm^3 after recalibration of the six model forms. De Vos (2005) also identified common model forms in PTFs, and recalibrated multiple PTFs. They reported the lowest standard deviation prediction error obtained after recalibration was 0.16 Mg/m^3. Khodaverdiloo et al. (2022) obtained RMSE values of 0.10 to 0.13 Mg/m^3 after recalibrating four PTFs. This study also identified common model forms amongst published PTFs, building on the categories developed by Nanko et al. (2014) and De Vos et al. (2005), categorizing PTFs into eight model types. The results after recalibration were affected by the dataset used, and the model type. The lowest RMSE obtained was 0.19 g/cm^3, and the highest CCC value was 0.68; which is comparable to the results produced by previous studies. This showed that recalibration is a method which improved the accuracy of existing PTFs for a wide range

of PTF model types, beyond the often employed models based solely on OC; and it is applicable to different regions for which no PTFs have been developed.

Chapter 3 investigated using machine learning to estimate missing values in legacy datasets. While other studies have compared the accuracy of different models when estimating soil attributes (Gharahi Ghehi et al., 2012; Gunarathna et al., 2019), this study faced the challenge of estimating missing values, but with a patchwork of available attributes with which to do so. The target attribute for this chapter was also BD, as it was present in less than 1.5% of horizons, and because it is required to calculate soil carbon stocks. However, other soil attributes in the dataset were also incomplete; for example, sand, silt and clay percentages were only available for ~10% of horizons, CEC and TN for ~14%, OC for 17%, and pH for 22%. To increase the number of horizons for which an estimated BD value could be made, other soil attributes with higher representation in the dataset were incorporated: depth, CF, order, PM, and textural class.

Using all 11 available attributes, every possible combination of those attributes was determined. The result was 513 models for predicting BD, and each was tested using the machine learner RF. Only one other similar study was identified, Benke et al. (2020), who tested the accuracy of 560 models for prediction of OC and electrical conductivity (EC). The models were ranked based on mean square prediction error (MSPE), and the most important predictors for OC and EC were identified. This study used CCC as the basis for ranking the models, and the results ranged from 0.92 to 0.51. Many of the models provided predictions for the same horizons, so the prediction with the highest accuracy was chosen. Models which incorporated continuous variables such as OC, CEC, pH, sand, clay, and TN had the highest accuracy; the best performing model was *BD = f(depth + CF + OC + pH + sand + clay).* However, as the continuous variables also had limited coverage, models which used the categorical variables provided the greatest number of predictions. The model which estimated ~ 63% of the missing BD values was *BD = f(depth + CF + order + PM + textural class);* this model still had a strong CCC value, at 0.78.

When the accuracy of the recalibrated equation-based PTFs was compared to the accuracy of the machine learning generated PTFs, the latter showed higher accuracy. The recalibrated PTFs were an improvement over the unrecalibrated PTFs, for the three datasets on which recalibration through NLS was tested; some PTFs showed

very large improvements while others were minimal, but overall recalibration had a positive effect on the accuracy of the PTFs. The accuracy depended both on the model form of the equation, and on the dataset used for recalibration. The PTFs developed with the RF algorithm also varied in their accuracy; 513 models were produced with varying combinations of input variables, and the resulting CCC values showed a range from 0.51 to 0.92. This range exceeded that of the recalibrated PTFs; while there were outliers in every dataset, the majority of the CCC values for those PTFs recalibrated on the Ontario (All Variables) dataset ranged from 0.63 to 0.68; for the BC (All Variables) the majority ranged from 0.40 to 0.58, and for the BC (Carbon and Bulk Density), the majority of the CCC values ranged from 0.36 to 0.55.

For both approaches to estimating missing values using PTFs that were investigated in this research, uncertainty estimates were also generated through QR. For both approaches, uncertainty was affected by both dataset size and predictor variables. Recalibration was carried out on three different datasets: the largest was from Ontario, and had 3,424 horizons with 8 variables; the mid-sized dataset was from BC, contained 1,199 horizons, and 2 variables; and the smallest was also from BC, had 396 horizons, with 8 variables. The PICP vs CL graphs showed that the uncertainty was overestimated by significant amount for the BC (All Variables) dataset, moderately overestimated for the BC (C and BD) dataset, and most accurately estimated for the Ontario (All Variables) dataset. The MPI graphs showed that uncertainty varied both by dataset used and by model form; uncertainty was greatest for the smallest dataset, and least for the largest dataset. The models with the lowest uncertainty overlapped with those with the highest accuracy; these were models with simple forms and OC or OM as the predictor.

When RF was used to test the 513 models, each model had a different sized training dataset depending on their combination of attributes, ranging from 191 to 1,450 horizons. Models with larger datasets showed through their PICP vs CL graphs that their uncertainty was more accurately estimated. Predictors played a secondary role in uncertainty, and models with less accurate predictors such as textural class and PM had higher uncertainty demonstrated on their MPI graphs. Somarathna et al. (2017) showed similar results when they tested the effect of dataset size on accuracy and uncertainty of soil OC predictions; as dataset size increased, accuracy increased, and uncertainty decreased. In this study, dataset size did not affect the accuracy of the machine learning results, but it did affect the accuracy of the PTFs which were recalibrated.

4.2. Limitations and Future Research

Using legacy data comes with limitations, the foremost being the limited amount of data. Small dataset sizes have shown to reduce the accuracy and increase the uncertainty of modeling predictions (Wu et al., 2022; Somarathna et al., 2017). To increase the size of a dataset, existing data can be augmented with new samples, where the locations of existing samples are fitted into a sampling design, and new sampling locations identified (Zhang et al., 2016). This approach maximizes the use of legacy data and while adding new data in the optimal locations.

The measurement of the target variable, in this case BD, is especially important, as these values are used to train every model used to predict missing values. In Chapter 2, a maximum of 1450 horizons had BD values, and the models used these values to then estimate BD for approximately 100,000 horizons. BD values are used in many calculations of other soil properties, such as porosity, volumetric moisture content, thermal conductivity, volumetric heat capacity, and penetration resistance (Al-Shammary et al., 2018). BD can be measured through multiple methods, including the excavation, clod and core methods (Throop et al., 2012). Using the core method may result in high bulk density values due to compaction when the core is pressed into the soil (Page-Dumroese et al., 1999). Throop et al. (2012) found that while the core method is the most often used way to determine BD, there are issues with the method such as coarse fragments too large to fit into the core and compaction due to core insertion. Al-Shammary et al. (2018) reviewed the literature on soil bulk density sampling methods and found that the accuracy of the core method depended on sampling depth, the size of the core, and the experience level of the sampler. They also noted that soil texture affected the accuracy of the excavation method and concluded that the radiation method had the highest accuracy but was costly and accuracy declined with depth. There are also different ways that BD can be measured, which depend on whether the coarse fragment material greater than 2 mm in diameter is included (Throop et al., 2012). All these variations introduce uncertainty into the BD values, which can then be propagated using PTFs to other values which are calculated using BD.

Another limitation to accurate predictions is the quality and availability of the input variables. Some variables, such as CF, have been found to be difficult to estimate, and their effect on soil can be variable (Holmes et al., 2021). There have been PTFs

developed on stony soils (Curtis and Post, 1964), but sampling for bulk density can be difficult in rocky soil (Frasier and Keiser, 1993). There are methods that were created to better sample soil with high stone or gravel content, such as filling a hole with paraffin wax consecutively in layers as it is excavated (Frasier and Keiser, 1993). Page-Dumroese et al. (1999) compared five methods of sampling bulk density in rocky soil, and found the results to be variable, with each method having both advantages and disadvantages. CF is required to calculate the fine fraction bulk density from the whole soil bulk density, and to calculate soil carbon stocks. The way in which BD measurements handle CF can affect the BD value by up to 26% (Throop et al., 2012).

Sand, silt and clay have been found to be good predictors for bulk density (Heinonen, 1977; Qiao et al., 2019; Tomasella and Hodnett, 1998) especially clay in low carbon soils (Dexter et al., 2008). This study showed mixed results; for recalibrated, equation-based PTFs, OC dominated as the most important predictor; however, when RF was used, sand and clay were present as attributes in the most accurate model and 28 of the top 40 models. As particle size fractions can be important predictors, it was assumed that textural class would also be a good predictor for BD. However, models which included textural class had lower accuracy and higher uncertainty that models which included sand and clay percentages. This is likely due to the accuracy of hand texture assessment in the field. For example, Salley et al. (2018) found that professional soil scientists had a 66% accuracy rate, while that of seasonal field technicians ranged from 27 to 41% accuracy, when compared to lab-determined measurements of textural class. While the results of laboratory methods of determining particle size distributions, such as sieving, hydrometer, pipette and laser diffraction, may vary with method used, and there is no universal method to which other methods can be compared (Eshel et al., 2004), hand texturing has been shown to still be less accurate than laboratory methods (Levine et al., 1989). Salley et al. (2018) suggested that hand texturing in the field could be improved through more training, providing a locally developed range of calibration soil samples for practice, and decision support tools such as mobile apps. Textural class was included in the model which predicted 63% of the missing BD values in this study; hence, an increase in its accuracy would have a large effect on model results.

A soil attribute that was found to be the most important predictor for BD in Chapter 2, and the most often included variable of the top 40 models in Chapter 3, was OC. OC is often determined through the measurement of organic matter (OM), such as

by loss-on-ignition (LOI) (Ball, 1964), then calculated using a conversion factor. Pribyl (2010) investigated this issue, compiling results from the literature which showed that the OC:OM ratio is not static, and on average is closer to 2 than 1.724. Factors which affect the OC:OM ratio include the method used to determine OM; the age of the soil; and depth (Pribyl, 2010). Périé and Ouimet (2008) found that their study of forest soils in Quebec was consistent with previous research that showed the OC:OM ratio changing with depth. Including depth as a covariate may reduce the effect of the variability of the OC:OM ratio when used in a machine learning model, and depth was included in every model in this study. However, it is a covariate that is not often used in equation-based PTFs. If a PTF is being developed through regression or recalibrated, it would be useful to determine an OC:OM ratio by depth before using the PTF.

Potential ways of improving the accuracy of PTF estimations include testing a variety of models and incorporating environmental variables. There are numerous machine learning algorithms which could be tested; for example, Shiri et al. (2017) compared heuristic gene expression programming (GEP), neural networks, RF, support vector machine, and boosted regression trees, and found the heuristic GEP model to have the highest performance. Artificial neural networks (ANNs) have produced good results in multiple studies (Al-Qinna and Jaber, 2013; Alaboz et al., 2021) as has k-nearest neighbour (kNN) (Botula et al., 2015, Gharahi Ghehi et al., 2012). Incorporating environmental covariates has also been shown to improve accuracy (Schillaci et al., 2021) and although this study focussed on using only soil attributes for BD prediction, it would be beneficial to explore including environmental covariates in future PTF studies, due to the limited availability of soil attribute values in the legacy dataset.

The two methods of improving the accuracy of PTFs which were explored in this thesis, as well as the approach to quantifying uncertainty which was coupled with both methods, showed positive results. These methods can be applied to other legacy soil datasets, so that the valuable data which they contain can be used to the greatest extent possible. With the collection of new data being frequently infeasible, the use of legacy data is often the only option; further, the significant resources which were put into data collection previously should not go to waste. Therefore, improving the accuracy of PTFs allows both the utilization of legacy data and provides better quality data for future DSM projects.

4.3. References

Abdelbaki, A.M. 2018. Evaluation of pedotransfer functions for predicting soil bulk density for U.S. soils. Ain Shams Engineering Journal, **9:** 1611-1619.

Alaboz, P., Demir, S., Dengiz, O. 2021. Assessment of various pedotransfer functions for the prediction of the dry bulk density of cultivated soils in a semarid environment. Communications in Soil Science and Plant Analysis, **52(7):** 724-742.

Al-Qinna, M.I., Jaber, S.M. 2013. Predicting soil bulk density using advanced pedotransfer functions in an arid environment. Transactions of the ASABE, **56(6):** 963-976.

Al-Shammary, A.A.G., Kouzani, A.Z., Kaynak, A., Khoo, S.Y., Norton, M. and Gates, W. 2018. Soil bulk density estimation methods: A review. Pedosphere, **28(4):** 581-596.

Ball, D.F. 1964. Loss-on-ignition as an estimate of organic matter and organic carbon in non-calcareous soils. J. Soil Sci., **15:** 84-92.

Benke, K.K., Norng, S., Robinson, N.J., Chia, K., Rees, D.B., Hopley, J. 2020. Development of pedotransfer functions by machine learning for prediction of soil electrical conductivity and organic carbon content. Geoderma, **366:** 114210.

Boschi, R.S., Bocca, F.F., Lopes-Assad, M.L.R.C., Assad, E.D. 2018. How accurate are pedotransfer functions for bulk density for Brazilian soils? Scientia Agricola, **75(1):** 70-78.

Botula, Y-D., Nemes, A., Van Ranst, E., Mafuka, P., De Pue, J., and Cornelis, W.M. 2015. Hierarchical pedotransfer functions to predict bulk density of highly weathered soils in Central Africa. Soil Sci. Soc. Am. J. **79:** 476-486.

Casanova, M., Tapia, E., Seguel, O., Salazar, O. 2016. Direct measurement and prediction of bulk density on alluvial soils of central Chile. Chilean Journal of Agricultural Research, **76(1):** 105-113.

Chen, S., Richer-de-Forges, A.C., Saby, N.P.A., Martin, M.P., Walter, C., Arrouays, D. 2018. Building a pedotransfer function for soil bulk density on regional dataset and testing its validity over a larger area. Geoderma, **312:** 52-63.

Curtis, R.O. and Post, B.W. 1964. Estimating bulk density from organic-matter content in some Vermont forest soils. Soil Sci. Soc. Am. J., **28(2):** 285-286.

De Vos, B., Van Meirvenne, M., Quataert, P., Deckers, J., and Muys, B. 2005. Predictive quality of pedotransfer functions for estimating bulk density of forest soils. Soil Sci. Soc. Am. J., **69:** 500-510.

Dexter, A.R., Richard, G., Arrouays, D., Czyż, E.A., Jolivet, C., Duval, O. 2008. Complexed organic matter controls soil physical properties. Geoderma, **144:** 620-627.

Eshel, G., Levy, G.J., Mingelgrin, U., and Singer, M.J. 2004. Critical evaluation of the use of laser diffraction for particle-size distribution analysis. Soil Sci. Soc. Am. J., **68:** 736-743.

Frasier, G.W. and Keiser, J. 1993. Thin layer measurement of soil bulk density. Journal of Range Management, **46(1):** 91-93.

Gharahi Ghehi, N., Nemes, A., Verdoodt, A., Van Ranst, E., Cornelis, W.M., Boeckx, P. 2012. Nonparametric techniques for predicting soil bulk density of tropical rainforest topsoils in Rwanda. Soil Sci. Soc. Am. J., **76:** 1172-1183.

Gunarathna, M.H.J.P., Sakai, K., Nakandakari, T., Momii, K., Kumari, M.K.N. 2019. Machine learning approaches to develop pedotransfer functions for tropical Sri Lankan soils. Water, **11(1940):** 1-23.

Han, G.-Z., Zhang, G.-L., Gong, Z.-T., and Wang, G.-F. 2012. Pedotransfer functions for estimating soil bulk density in China. Soil Science, **177(3):** 158-164.

Heinonen, R., 1977. Towards “normal” soil bulk density. Soil Sci. Soc. Am. J. **41(6):** 1214–1215.

Holmes, K.W., Griffin, E.A., van Gool, D. 2021. Digital soil mapping of coarse fragments in southwest Australia: targeting simple features yields detailed maps. Geoderma, **404:** 115282.

Kaur, R., Kumar, S., and Gurung, H.P. 2002. A pedo-transfer function (PTF) for estimating soil bulk density from basic soil data and its comparison with existing PTFs. Aust. J. Soil Res., **40:** 847-857.

Khodaverdiloo, H., Bahrami, A., Rahmati, M., Vereecken, H., Miryaghoubzadeh, M., Thompson, S. 2022. Recalibration of existing pedotransfer functions to estimate soil bulk density at a regional scale. Eur. J. Soil Sci., **73:** e13244.

Levine, S.J., Post, D.F., Ellsworth, T.J. 1989. An evaluation of student proficiency in field estimation of soil texture. J. Agron. Educ., **18:** 100-104.

Nanko, K., Ugawa, S., Hashimoto, S., Imaya, A., Kob ayashi, M., Sakai, H., Ishizuka, S., Miura, S., Tanaka, N., Takahashi, M., Kaneko, S. 2014. A pedotransfer function for estimating bulk density of forest soil in Japan affected by volcanic ash. Geoderma, **213:** 36-45.

Nasta, P., Palladino, M., Sica, B., Pizzolante, A., Trifuoggi, M., Toscanesi, M., Giarra, A., D’Auria, J., Nicodemo, F., Mazzitelli, C., Lazzaro, U., Di Fiore, P., Romano, N. 2020. Evaluating pedotransfer functions for predicting soil bulk density using

hierarchical mapping information in Campania, Italy. Geoderma Regional, **21:** e00267.

Page-Dumroese, D.S., Jurgensen, M.F., Brown, R.E., and Mroz, G.D. 1999. Comparison of methods for determining bulk densities of rocky forest soils. Soil Sci. Soc. Am. J., **63:** 379-383.

Périé, C., and Ouimet, R. 2008. Organic carbon, organic matter and bulk density relationships in boreal forest soils. Canadian Journal of Soil Science, **88(3):** 315-325.

Pribyl, D.W. 2010. A critical review of the conventional SOC to SOM conversion factor. Geoderma, **156:** 75-83.

Qiao, J., Zhu, Y., Jia, X., Huang, L., Shao, M., 2019. Development of pedotransfer functions for predicting the bulk density in the critical zone on the Loess Plateau, China. J. Soil. Sediment., **19:** 366–372.

Reidy, B., Simo, I., Sills, P. and Creamer, R.E. 2016. Pedotransfer functions for Irish soils – estimation of bulk density (ρb) per horizon type. SOIL, **2:** 25-39.

Salley, S.W., Herrick, J.E., Holmes, C.V., Karl, J.W., Levi, M.R., McCord, S.E., van der Waal, C., Van Zee, J.W. 2018. A comparison of soil texture-by-feel estimates: implications for the citizen soil scientist. Soil Sci. Soc. Am. J., **82:** 1526-1537.

Schillaci, C., Perego, A., Valkama, E., Märker, M., Saia, S., Veronesi, F., Lipani, A., Lombardo, L., Tadiello, T., Gamper, H.A., Tedone, L., Moss, C., Pareja-Serrano, E., Amato, G., Kühl, K., Dămătîrcă, C., Cogato, A., Mzid, N., Eeswaran, R., Rabelo, M., Sperandio, G., Bosino, A., Bufalini, M., Tunçay, T., Ding, J., Fiorentini, M., Tiscornia, G., Conradt, S., Botta, M., Acutis, M. 2021. New pedotransfer approaches to predict soil bulk density using WoSIS soil data and environmental covariates in Mediterranean agro-ecosystems. Science of the Total Environment, **780:** 146609.

Sevastas, S., Gasparatos, D., Botsis, D., Siarkos, I., Diamantaras, K.I. and Bilas, G. 2018. Predicting bulk density using pedotransfer functions for soils in the Upper Anthemountas basin, Greece. Geoderma Regional, **14:** e00169.

Shiri, J., Keshavari, A., Kisi, O., Karimi, S., Iturraran-Viveros, U. 2017. Modeling soil bulk density through a complete data scanning procedure: heuristic alternatives. Journal of Hydrology, **549:** 592-602.

Somarathna, P.D.S.N., Minasny, B., Malone, B.P. 2017. More data or a better model? Figuring out what matters most for the spatial prediction of soil carbon. Soil Sci. Soc. Am. J., **81:** 1413-1426.

Taulya, G., Tenywa, M.M., Majaliwa, M.J.G., Odong, T.L., Kaingo, J., and Kakone, A. 2005. Validation of pedotransfer functions for soil bulk density estimation on a

Lake Victoria Basin soilscape. African Crop Science Conference Proceedings, **7:** 1049-1056.

Throop, H.L., Archer, S.R., Monger, H.C., Waltman, S. 2012. When bulk density methods matter: Implications for estimating soil organic carbon pools in rocky soils. Journal of Arid Environments, **77:** 66-71.

Tomasella, J., Hodnett, M.G., 1998. Estimating soil water retention characteristics from limited data in Brazilian Amazonia. Soil Sci., **163(3):** 190–202.

Vasiliniuc, I. and Patriche, C.V. 2015. Validating soil bulk density pedotransfer functions using a Romanian dataset. Carpathian Journal of Earth and Environmental Sciences, **10(2):** 225-236.

Wu, T., Wu, Q., Zhuang, Q., Li, Y., Yao, Y., Zhang, L., and Xing, S. 2022. Optimal sample size for SOC content prediction for mapping using the random forest in cropland in northern Jiangsu, China. Eurasian Soil Science, **55(12):** 1689-1699.

Yi, X.S., Li, G.S., and Yin, Y.Y. 2016. Pedotransfer functions for estimating soil bulk density: A case study in the Three-River Headwater region of Qinghai Province, China. Pedosphere, **26(3):** 362-373.

Zhang, S.J., Zhu, A.X., Liu, J., Yang, L., Qin, C.Z., An, Y.M. 2016. An heuristic uncertainty directed field sampling design for digital soil mapping. Geoderma, **267:** 123-136.