frontiers | Frontiers in Environmental Science

TYPE Methods
PUBLISHED 24 January 2023
DOI 10.3389/fenvs.2023.1009191

OPEN ACCESS

EDITED BY
Juergen Pilz,
University of Klagenfurt, Austria

REVIEWED BY
Ana Horta,
Charles Sturt University, Australia
Rahim Alhamzawi,
University of Al-Qadisiyah, Iraq
B. Rajanarayan Prusty,
Alliance University, India

*CORRESPONDENCE
Ségolène Dega,
segolene.dega@ufz.de

SPECIALTY SECTION
This article was submitted to Environmental Informatics and Remote Sensing,
a section of the journal
Frontiers in Environmental Science

RECEIVED 01 August 2022
ACCEPTED 10 January 2023
PUBLISHED 24 January 2023

CITATION
Dega S, Dietrich P, Schrön M and Paasche H (2023), Probabilistic prediction by means of the propagation of response variable uncertainty through a Monte Carlo approach in regression random forest: Application to soil moisture regionalization.
Front. Environ. Sci. 11:1009191.
doi: 10.3389/fenvs.2023.1009191

Probabilistic prediction by means of the propagation of response variable uncertainty through a Monte Carlo approach in regression random forest: Application to soil moisture regionalization

Ségolène Dega[1]*, Peter Dietrich[1,2], Martin Schrön[1] and Hendrik Paasche[1]

[1]Department of Monitoring and Exploration Technologies, Helmholtz Centre for Environmental Research (UFZ), Helmholtz Association of German Research Centres (HZ), Leipzig, Germany, [2]Department of Environmental and Engineering Geophysics, University of Tübingen, Tübingen, Baden-Württemberg, Germany

Probabilistic predictions aim to produce a prediction interval with probabilities associated with each possible outcome instead of a single value for each outcome. In multiple regression problems, this can be achieved by propagating the known uncertainties in data of the response variables through a Monte Carlo approach. This paper presents an analysis of the impact of the training response variable uncertainty on the prediction uncertainties with the help of a comparison with probabilistic prediction obtained with quantile regression random forest. The result is an uncertainty quantification of the impact on the prediction. The approach is illustrated with the example of the probabilistic regionalization of soil moisture derived from cosmic-ray neutron sensing measurements, providing a regional-scale soil moisture map with data uncertainty quantification covering the Selke river catchment, eastern Germany.

KEYWORDS

uncertainty propagation, probabilistic prediction, Monte Carlo, quantile regression random forest, soil moisture

1 Introduction

Knowledge about state variables of the Earth is a primary source of information when striving to model the state and dynamics of processes in the Earth and environment. Among others, soil moisture (SM) has been identified as an essential climate variable when studying land surface ecosystems (Gruber and Peng, 2022) and is, therefore, a variable of particular interest. SM measurements for local- and regional-scale studies are still often measured sparsely (Schröter et al., 2015), i.e., by sampling a low number of measurements at distinct points distributed over the survey area or along a few trajectories in a larger area, leaving large gaps of information in the resultant dataset. In the presence of such a sparsely sampled dataset, one can resort to the theory of multiple regression problems (MRPs), which deals with finding a relationship between a response variable and other explanatory variables called predictors (Kuhn & Johnson, 2013). The found relationship can be used to estimate the value of the

response variable where it has not been measured. Unlike simple mathematical interpolation methods, e.g., nearest neighbor interpolation, MRPs allow incorporating information from other dense data comprising lower gaps between the sampling locations and are, therefore, more suited when the response variable data are sparse. As it is relatively common in geoscience to only have access to sparse spatial data (Howarth, 2001), MRPs are widely used by Earth scientists to derive spatially continuous maps, also in the context of SM prediction (Adab et al., 2020; Carranza et al., 2021), but they do not always include uncertainty quantification of their prediction results (Perez Dias et al., 2020). Schmidt et al. (2020) consider uncertainty associated with data as noise and characterize it as unwanted information that impairs correct information retrieval from data. Quantified uncertainty defines the knowledge about what is not known about the prediction results, and it is consequently of value for any application of the prediction made. Therefore, a soil map can only be considered complete if its uncertainties have been explicitly quantified (Heuvelink, 2014). It is important to distinguish between point uncertainty and spatial uncertainty. In the first case, uncertainties are computed only at some locations defined from a set of independent soil observations and are, in fact, a validation process of the model calibration. Lagacherie et al. (2019) showed that these point uncertainty metrics suffer variability depending on the training dataset size and are, therefore, themselves prone to uncertainty. They give information on how well the model is trained but cannot infer the uncertainty of any new prediction. Spatial uncertainty quantification, however, which is the focus of this work, aims at estimating uncertainty for every point of the predicted soil property, as, for example, performed by Heuvelink (2014) and Nauman and Duniway (2019).

McBratney et al. (2003) were one of the first to resort to MRPs in geoscience and proposed a soil spatial prediction function which predicts a soil attribute from other soil covariates with a spatially autocorrelated error. Nowadays, these problems often resort to machine learning algorithms since their level of required preconceptions about the shape of the regression model is maximally low. Lorenzetti et al. (2015) showed that machine learning, and more specifically, support vector machine, gives more reliable results than traditional pedology approaches to assess the class frequency in soil map legends. Regression trees (Breiman et al., 1984) and random forests (Breiman, 2001) are also commonly used to solve MRP in soil mapping as they are intuitive, computationally cheap, and can consider many explanatory variables. Nussbaum et al. (2018) showed that random forest (RF) performs slightly better than other statistical methods for the spatial assessment of a soil function in the case of a large set of predictors. RF was also successfully applied by Hengl et al. (2018) to predict spatial variables, including information derived from observation locations as the covariate; it was also applied by Poggio et al. (2021) to produce global maps of soil properties for SoilGrids (https://soilgrids.org/). Although previous references do not always include uncertainty quantification, one can observe an increasing awareness among Earth scientists of the necessity to quantify the uncertainties (Perez Dias et al., 2020; Paasche et al., 2021). McBratney et al. (2003) were one of the precursors in this aspect in the soil analysis community. They proposed a soil attribute mapping approach with a measure of the uncertainty as they claim that uncertainty is necessary information to handle a possible lack of training data or data with poor quality. The necessity to quantify uncertainty in soil property prediction was also earlier addressed by Heuvelink and Webster (2001), who reviewed three statistically based models allowing such a quantification. It is indeed a strength of statistical methods to allow having a measure of uncertainty in the prediction by comparison of inference models (Lagacherie, 2008).

Spatial uncertainty quantification can be addressed in different ways. When the uncertainty of the input data (observations and covariates) is unknown, one can resort to models which can provide prediction error estimation. Hengl et al. (2004) used regression kriging for spatial predictions of soil variables and were able to compute the variance of the prediction error to show the uncertainty together with the prediction map. When RF is used for predictions, many examples exist in the literature to quantify the uncertainty (Wager et al., 2014; Mentch & Hooker, 2016; Baake, 2018). Uncertainty quantification can also be performed with quantile regression random forests (Meinshausen, 2006). Quantile regression random forests (QRRF) not only provide the mean of the response variable such as conventional RF but also its full distribution. QRRF was, for example, successfully applied by Vaysse and Lagacherie (2017) to predict the uncertainties of digital soil-mapping products, leading to better predicted patterns of uncertainty than regression kriging. Poggio et al. (2021) used QRRF to routinely quantify the spatial uncertainty of digital soil mapping products of various state variables (e.g., accessible *via* https://soilgrids.org). On the other hand, if uncertainties of the input data are known and quantified, one can use this information with different approaches. van der Westhuizen et al. (2022) used the measurement uncertainty in digital soil mapping during the machine learning model calibration, giving less weight to measurements with high variance error. A similar approach was used by Wadoux et al. (2019) for the calibration of a convolutional neural network. These works show that under some conditions, this strategy can lead to greater accuracy of predictions. Another classical approach (Heuvelink, 1998) for the propagation of the uncertainty of the input data onto the prediction is the Monte Carlo (MC) approach—relying on the repetition of the whole prediction process with a different training dataset drawn from the input distribution at each run. In the case of non-Gaussian probability density distributions associated with the data, MC-related approaches are considered a suitable choice for realistic uncertainty propagation from input data through mathematical processing into the resultant output data (Durbin and Koopman, 1997; JCGM et al., 2008).

Among methods designed to quantify uncertainty in RF for the case of unknown or ignored data uncertainties associated with the response variable and predictors, QRRF seems to be the most used in the literature. Since QRRF takes the raw training database as input without any quantification of the training data uncertainties, the uncertainties provided in the output cannot reflect the noise in the input training data but rather an approximation of the regression model uncertainty expressed by the RF itself. In such an approach, the uncertainty quantification of the predicted data is reduced to a purely mathematical task ignoring the impact of uncertainty propagation which requires detailed knowledge of the nature of the input data and the processing procedure (JCGM et al., 2008). As far as our knowledge goes, uncertainty quantification for SM prediction has been performed with QRRF (Carranza et al., 2021) but not through the propagation of SM measurement uncertainty. Moreover, many references encountered in the literature for soil property prediction compare different methods for the prediction (machine learning, model-based) but not for uncertainty quantification. Therefore, the strengths and weaknesses of the models are well studied, whereas such comparisons for uncertainty quantification remain less documented.

In this study, the main goal is to demonstrate the applicability of a random forest-based Monte Carlo approach to quantify the uncertainty of SM prediction and determine how the uncertainty of the SM used to derive the regression model impacts the uncertainty of the SM prediction. As a benchmark, we use prediction uncertainty achieved by QRRF, which ignores the quantified uncertainty of the input data. Our response variable is gravimetric soil moisture sparsely measured by mobile cosmic-ray neutron sensing (McJannet et al., 2017; Schrön et al., 2018; Jakobi et al., 2020; Schrön et al., 2021). The second goal of this study is to perform a comparative analysis of QRRF and MC approaches to understand the kind of uncertainty quantified by each method.

This paper starts by describing the mathematical background of uncertainty in MRP and introduces various terms to refer to the different kinds of uncertainty which can be found in MRP. Then, the study area and the input data, the sparsely measured gravimetric SM, and the predictors are described. The paper then gives details of the methods used for the prediction, RF, uncertainty quantification, QRRF, and MC. Similarities and discrepancies in the output distributions between both methods are identified by recognizing areas of low and high predicted uncertainties and comparing them with uncertainties of the input data. Finally, a unified map of the SM prediction is provided together with the quantified uncertainties.

2 Uncertainty in multiple regression problems

Let $\boldsymbol{x_a}$ with $\boldsymbol{x_a} = (x_{a1}, x_{a2}, \ldots, x_{ak})$ be a one-dimensional variable with k numeric values representing different states of a quantity over time or space. For example, $\boldsymbol{x_a}$ could be a map comprising k numeric values of a quantity, such as topographic elevation. A function $f(\boldsymbol{x_a})$ could describe the linkage between $\boldsymbol{x_a}$ and a one-dimensional variable $\boldsymbol{y}$. In a problem of the type $\mathbf{y} = f(\boldsymbol{x_a})$, $\boldsymbol{y}$ can be named the regressand, the response variable, or the dependent variable, whereas $\boldsymbol{x_a}$ is referred to as the regressor, a feature, an independent variable, a covariate, an explanatory variable, or a predictor. In the following, we will refer to $\boldsymbol{y}$ and $\boldsymbol{x_a}$ as the response variable and a predictor, respectively. If the response variable depends on more than one predictor, we can express the link between all predictors and $\boldsymbol{y}$ by

$$\boldsymbol{y} = f(\boldsymbol{X}), \tag{1}$$

with $\boldsymbol{X}$ being a multi-dimensional variable $\boldsymbol{X} = (\boldsymbol{x_a}, \boldsymbol{x_b}, \ldots)$. The function f can be a regression model and can be parametric or non-parametric. If f is known, the response variable can be predicted at every instance in space or time when the predictors are known.

In practical problems, $\boldsymbol{X}$ is regarded as known at all time or space instances of interest, but f is usually not known *a priori*. Instead, $\boldsymbol{y}$ is known at some instances of known $\boldsymbol{X}$, which allows deriving f. If some expectation about the form of f exists (e.g., linearity), parametric or semi-parametric modeling could be used. If no reliable assumption about the form of f can be formulated, e.g., as an analytical function and its parameters, the regression model has to be learned and described non-parametrically, e.g., by using regression trees.

Equation 1 assumes a perfect fit between the response variable and the predictors, but in practice, a multi-dimensional disturbance or noise variable $\boldsymbol{U}$ is present:

$$\boldsymbol{y} = f(\boldsymbol{X}, \boldsymbol{U}). \tag{2}$$

$\boldsymbol{U}$ can include random or systematic errors, also known as uncertainty about precision and accuracy, respectively, reducing the accuracy and precision of the predictors and the known instances of $\boldsymbol{y}$. Solving the regression problem suffers from uncertainty propagation from $\boldsymbol{X}$ and $\boldsymbol{y}$ into f. Additionally, uncertainties linked to the realistic definition of the character of f, e.g., assumed linearity, will overlap with the uncertainties from $\boldsymbol{X}$ and $\boldsymbol{y}$. Hence, $\boldsymbol{U}$ is an aggregate of uncertainties originating from $\boldsymbol{X}$, $\boldsymbol{y}$, and f.

If the information stored in $\boldsymbol{X}$ and $\boldsymbol{y}$ is altered by errors due to limited precision and accuracy, we cannot retrieve it without uncertainty. Since errors in $\boldsymbol{X}$ and $\boldsymbol{y}$ propagate into the finding of f, uncertainty never diminishes and propagates through any processing step of uncertain data making the finding of f an indeterminate task. In a deterministic regression problem, the uncertainty of f rooted in the choices of the definition of f is ignored. Using a determined f to predict $\boldsymbol{y}$ from $\boldsymbol{X}$ will give determined predictions. In a stochastic regression problem, we face the presence of a non-zero disturbance variable $\boldsymbol{U}$ resulting in an indeterminate finding of f and indeterminate predictions of $\boldsymbol{y}$ when applying f to uncertain $\boldsymbol{X}$.

Uncertainty measures the level of imperfect information coded in data. Throughout this study, only the response variable uncertainty is considered, and the predictor uncertainty and uncertainty added by the model itself or the selection of the modeling method are not considered. In the case of the response variable, we must distinguish between the uncertainty related to data acquired for solving the regression problem and the uncertainty associated with the predicted response variable. For finding f, we need co-located instances of predictor and response variables. These measured instances of $\boldsymbol{y}$ are referred to as the training response variable, whereas predicted instances of the response variable are referred to as the predicted response variable. Data uncertainty is not always equivalent to measurement uncertainty. Data can be derived from measurements by applying a processing function g to the latter. Uncertainties in the definition of g or a generally methodologically inappropriate choice of g, for example, in order to close spatial or temporal knowledge gaps in the measurements, overlap with the measurement uncertainty.

As the response variable, we use gravimetric soil moisture, which is known in some instances, so that we can build a regression model f linking $\boldsymbol{y}$ with $\boldsymbol{X}$. The measured quantity is the counting number of neutrons per time unit (Zreda et al., 2012). This quantity has to be transformed by a function g into gravimetric soil moisture and associated uncertainties, while the measurement uncertainty of neutron counts has been already propagated through g (Jakobi et al., 2020; Schrön et al., 2021).

We solve our regression problem non-parametrically by using an RF to build f from available and co-located instances of $\boldsymbol{X}$ and $\boldsymbol{y}$. Theoretically, this leaves our approach free of preconceptions (e.g., linearity) about the character of f. However, random forests require the definition of some parameters, e.g., the number of trees or the depths of trees. Often, no hard rules exist for the "right" selection of the necessary parameter settings. So, here, uncertainty in setting the "right" model parameters must be expected to overlap with uncertainty from $\boldsymbol{X}$ and $\boldsymbol{y}$ when building an RF regression model. In addition, RFs are inherently biased when it comes to the utilization of the built regression model for prediction. Since random forests do not incorporate the concept of extrapolation, all predictions of $\boldsymbol{y}$ will always be included in the range of the instances of $\boldsymbol{y}$ used for learning the regression model (Zhang et al., 2017). Such methodological limitations may inherently bias prediction outcomes.

In this study, we focus on $\boldsymbol{y}$ uncertainty, and we will differentiate between training response variable uncertainties and predicted response

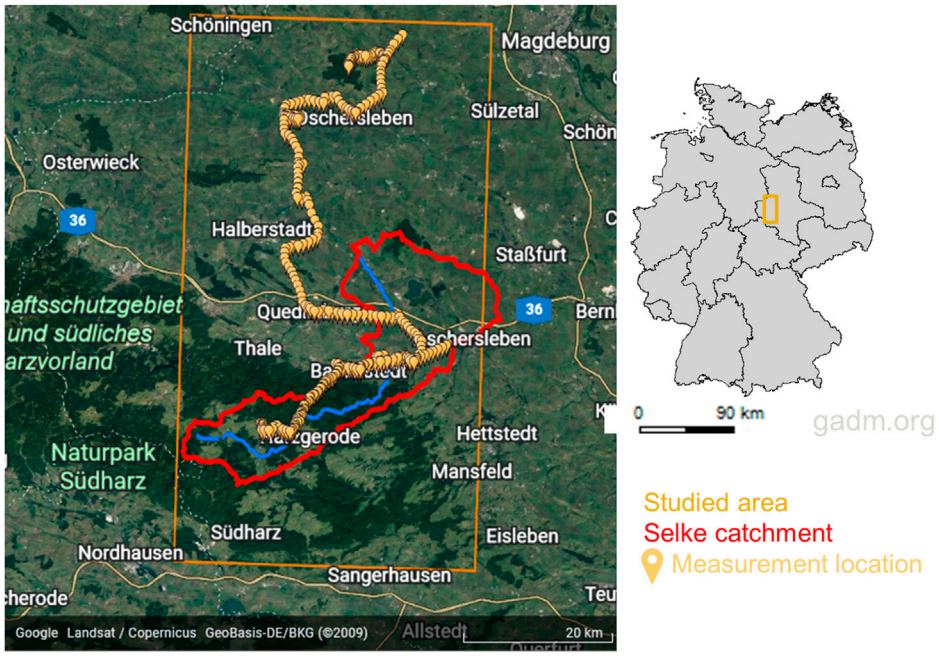

FIGURE 1
Map of the mobile CRNS measurement campaign overlapping the Selke catchment in red and the rectangular studied area for the probabilistic prediction in orange and the global location within Germany. Background map is from Google Earth, earth.google.com/web/.

variable uncertainties. Data uncertainty may contain aspects resulting from measurement uncertainty, methodological choices, and application of uncertain functions to produce processed instances of data. The uncertainty of the regression model f is referred to as model uncertainty. Since the regression model is derived from $\boldsymbol{X}$ and $\boldsymbol{y}$ and their uncertainty, the differentiation between data and model uncertainty appears a little fuzzy. In the following analyses, we focus on the different kinds of response variable uncertainty.

3 The database

3.1 The response variable gravimetric soil moisture

The study area covers the Selke catchment located in the Harz Mountains and the Harz foreland of the federal state of Saxony-Anhalt, Germany. The area is part of the hydrological TERENO Harz/Central German Lowland Observatory (Zacharias et al., 2011, see also Figure 1). The catchment has high gradients of landscape regarding elevation and other soil properties and has been extensively studied (Sinha et al., 2016; Wollschläger et al., 2016; Yang et al., 2018; Yang et al., 2019; Winter et al., 2021). The northern sub-catchment has an elevation of around 100 m and is mainly covered by agricultural areas, while most of the southern sub-catchment, with a higher elevation (up to 590 m), is covered by dense forest.

We used data from a mobile cosmic-ray rover survey conducted on 3 September 2015, conducted by Kasner (2016), comprising 314 samples. These measurements have been made with a mobile CRNS measurement system mounted on a car driving approximately 150 km through the Selke catchment along the existing public road network. Driving speed was kept constant as much as possible without endangering public traffic. The dataset might be considered small, making uncertainty quantification more challenging (Lagacherie et al., 2019). Another challenge associated with mobile CRNS is that data are collected along the road, as mobile sensors are mounted on a car, so the area of interest for the prediction is not homogenously covered by measurements as it would be with a regular grid over the area.

For the presented study, we focus on the spatial prediction of gravimetric soil moisture by means of static predictors (i.e., which stay constant for the considered day) since we consider only measurements from a campaign conducted within 7 h on a cloudy day without precipitation during the campaign.

The measured soil moisture is derived from neutron counts N per time unit at the position of the CRNS sensor. The measured data have been processed in order to obtain a corrected neutron count N^* to be converted into gravimetric soil water content:

$$N^* = C_{neutron} * C_{pressure} * C_{humidity} * C_{road} * C_{veg} * N. \quad (3)$$

$C_{pressure}$ and $C_{humidity}$ are standard corrections for air pressure and air humidity (Hawdon et al., 2014) as air and water molecules attenuate incoming neutron rates. These factors are computed from temperature, pressure, and humidity data measured by sensors mounted on the CRNS rover. $C_{neutron}$ corrects for incoming neutron radiation which varies due to solar activity (Schrön et al., 2015) and is derived from the Jungfraujoch neutron monitor. Schrön et al. (2018) showed how roads introduce a significant bias in neutron counting, leading to an overestimation of the final soil moisture values; hence, the need for a correction of C_{road}. The correction factor was derived from information on road type from the OpenStreetMap (www.openstreetmap.org) road network data. In addition, all data in urban areas were removed based on CORINE land use classification data. Finally, biomass water impacts the measurement without being related to soil moisture (Baroni et al., 2018). That is why a biomass correction C_{veg} also based on CORINE

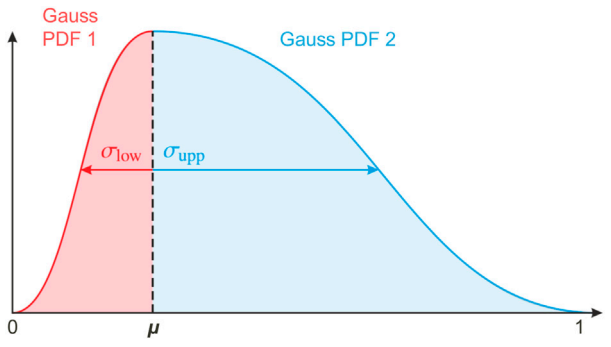

FIGURE 2
Asymmetrical Gaussian distribution used for soil moisture uncertainties.

land use data was applied, which consists in decreasing, by 5%, the neutrons count in forests to compensate for the high hydrogen concentration rate in these areas.

The uncertainty of the measurement originates from various sources. The main one is the stochastic uncertainty of neutron detection. Zreda et al. (2012) showed that neutron counting uncertainty has a Gaussian distribution whose standard deviation is defined by $\sigma_N = \sqrt{N}$ for N counted neutrons.

Moreover, the soil moisture heterogeneity in the footprint, which is particularly wide for high driving speeds or long aggregation periods, brings more variability in the neutron count. In this study, data were recorded every minute, and a rolling average over a window of three minutes was applied to reduce the stochastic error. With driving speeds ranging between 20 and 60 km/h, this corresponds to footprint lengths of 1–3 km. This uncertainty, however, could not be precisely quantified.

Gravimetric soil moisture is obtained from the corrected neutron count with the approach used by Köhli et al. (2021), shown in Eq. 4, which is equivalent to the relationship described by Desilets et al. (2010):

$$\theta(N^*) \approx p_0 \frac{1 - N^*/N_{max}}{p_1 - N^*/N_{max}}. \quad (4)$$

where N_{max} is the maximum neutron flux under dry conditions, and p_i represents calibration parameters. Following Schrön et al. (2021), the symmetrical uncertainty $\sigma_N{}^*$ from the corrected neutron count rate has been converted to soil water content uncertainty $\sigma_{\theta\pm}$, which is asymmetric due to the non-linearity of the conversion function:

$$\sigma_{\theta\pm} = \theta(N^*) - \theta\left(N^* \pm \sigma_N^*\right). \quad (5)$$

This error distribution has been approximated by a distribution modeled by merging the lower and upper half of two Gaussian distributions at the mean μ and (σ_{low}, σ_{upp}) as left and right standard deviations (Figure 2). Figure 3 shows all the measurements selected for this study with the quantified uncertainties expressed by the 5, 25, 75, and 95 quantiles.

3.2 The predictors

The predictors used for the soil moisture predictions are shown in Figure 4. They were chosen to be datasets freely available at the German national level and are known to have an impact on the soil moisture. They were all re-gridded to a 250-m grid to match the spatial resolution of the measured CRNS data used as the response variable.

The choice of the covariates is of great importance as they should be able to explain the spatial variability of soil moisture as much as possible. Topography's impact on soil moisture has been widely studied and proved to be a determinant factor of soil moisture variations, especially under wet conditions (Western et al., 1999a; Schröter et al., 2015). Topographic information can be made of various variables; for this study, we keep the elevation, slope, and aspect information. However, these variables do not explain all the soil moisture variations. Western et al. (2004) showed that depending on the catchment, soil properties and rainfall can explain better the spatial variability than topography. The same observation was made by Korres et al. (2010), and soil properties such as clay content, sand content, density, soil organic carbon percentage, and precipitation heights were identified to have an impact on the soil moisture. Moreover, soil temperature is also known to be related to soil moisture (Lakshmi et al., 2003) and was also incorporated as a covariate in our study.

The elevation has been derived from the DGM50 dataset provided by the German Federal Agency for Cartography and Geodesy. The DGM50 is a German topography model with a 50-m spatial resolution (http://www.bkg.bund.de). The elevation data were used to compute the topographic slope and aspect maps. The bulk density, clay and sand proportion, and soil organic carbon datasets were obtained from the SoilGrids database (https://soilgrids.org/, accessed on 01/14/2022) and came with a 250-m spatial resolution. These maps were derived from global soil profile information and covariate data, as described by Poggio et al. (2021).

Precipitation and soil temperature data for 2 days before the measurement campaign have been acquired from the German Weather Service (https://www.dwd.de, accessed on 01/24/2022), with a 1,000-m spatial resolution. Precipitation data were accumulated day-wise, whereas soil temperature data were provided as daily averages. Instead of considering the precipitation and soil temperature data on a daily basis, we computed the mean precipitation and soil temperature over 2 days before the campaign and used this aggregated information as predictor datasets in our MRP.

4 Methodology for solving the MRP

4.1 Decision trees

In this section, we will give an overview of the algorithms used for the predictions. This study uses RFs, introduced by Breiman (2001) and based on decision trees. A decision tree is a structure made of nodes which represent a "test" on a predictor, based on which the tree splits into branches leading to other nodes. The construction of the tree uses **n** independent observations (Y_i, X_i), $i = 1 \ldots n$. At each node, the algorithm looks at all predictors and their values and solves an optimization problem to decide which predictor leads to the more efficient split, i.e., those samples with a similar target are grouped together. This is performed by minimizing the mean squared error (MSE) between the actual output value and the one predicted across all candidate splits, determined by a predictor and a threshold value. The end node of the branch that does not split anymore is a leaf and gives an output of the regression task. Once the tree is constructed, the prediction $\hat{\mu}(x)$ for a new data point $X = x$ is a weighted average of the original observations Y_i, $i = 1 \ldots n$:

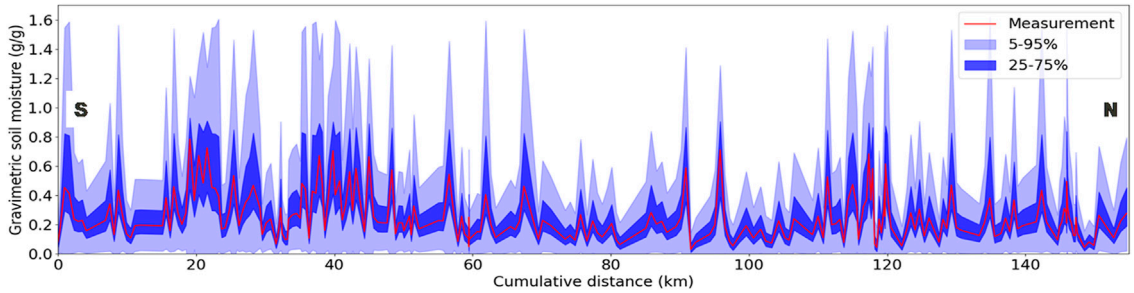

FIGURE 3
Soil moisture measurements with quantified uncertainties (fifth, 25th, 75th, and 95th percentiles) along driven distance.

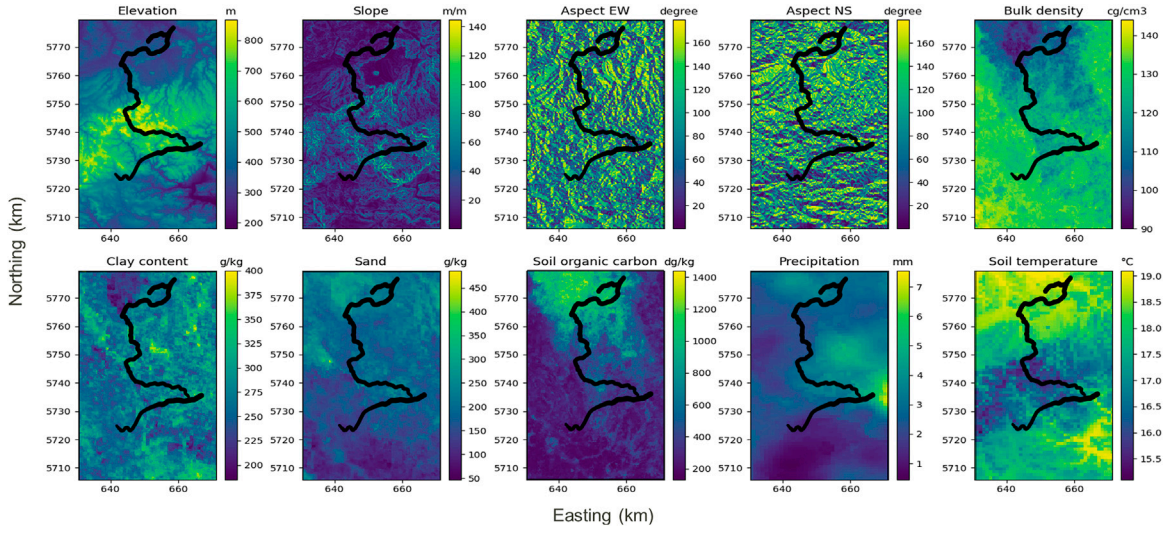

FIGURE 4
Features used in random forest for soil moisture determination.

$$\hat{\mu}(x) = \sum_{i=1}^{n} w_i(x,\theta) Y_i,$$

where $w_i(x,\theta)$ is the weight vector, and it is a positive value if the observation (Y_i, X_i) is part of the same leaf of the tree built from the random vector of variables θ in which x was dropped; otherwise, it equals 0. The weights add up to one.

4.2 Random forests

RFs are an ensemble learning method composed of multiple decision trees. The algorithm is based on the concept of bagging: one trains multiple trees with various subsamples of the training dataset and uses the average of all predictions. Some randomness is also included at each tree and each node when selecting a predictor to split on, and only a random subset of predictor variables is considered. This way, random forests are less dependent on the training dataset than individual trees to achieve a reduced variance and have better control on overfitting (Kuhn & Johnson, 2013). The prediction in RF is approximated by the averaged prediction of k single trees, and we can then construct $w_i(x)$ to be the average of $w_i(x,\theta)$ over this collection of trees:

$$w_i(x) = k^{-1} \sum_{t=1}^{n} w_i(x,\theta_t).$$

Finally, the prediction of RF for a new data point $X = x$ is given by

$$\hat{\mu}(x) = \sum_{i=1}^{n} w_i(x) Y_i.$$

For this work, we use the RandomForestRegressor tool from the Python library Sklearn. The most important parameters of the RF are the number of trees in the forest and the maximum depth of the tree. The latter represents the length of the longest path from the tree root to a leaf. Model tuning was performed prior to our study to find the optimal values of both these parameters. Different numbers of trees in the RF, from 20 to 200, were combined with different numbers of tree depths (between 3 and 17). For each combination, the MSE of the prediction was computed and is displayed in Figure 5. A total of 40 trees were kept as it results in a trough in all MSE curves, and the maximum depth was set to 8 to avoid

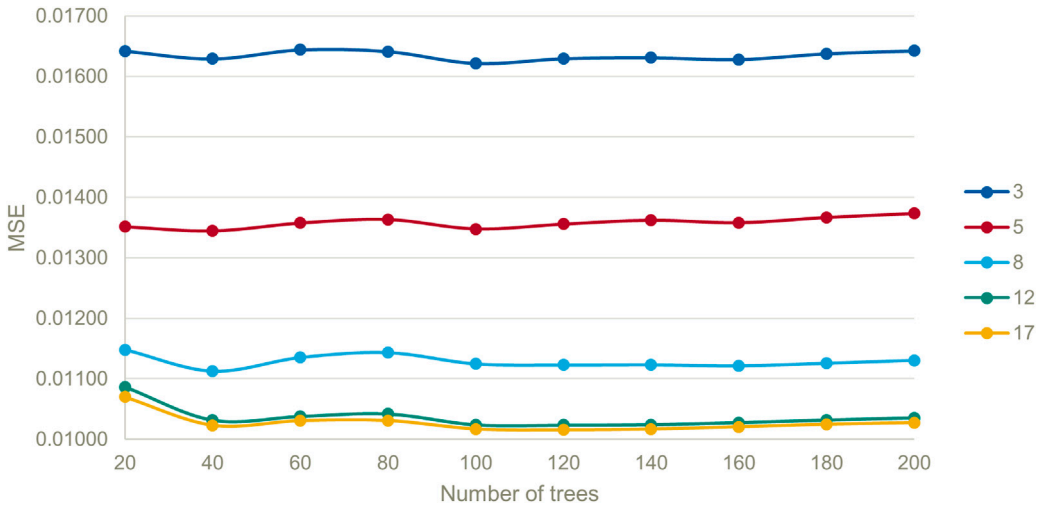

FIGURE 5
Comparison of the mean squared error (MSE) in RF training vs the number of trees in the forest for different forest depths. The number of 40 trees leads to a lower MSE for all depth configurations.

overfitting, although higher depths result in slightly lower MSE. With this choice, we ensure that each tree can comprise up to 128 leaves, which statistically ensures a moderate averaging effect when fitting the measures of the response variable in order to lower the risk of data overfitting when learning the regression model.

Feature importance analysis is performed using the tool provided for regression random forest in the Python Scikit-learn library. The importance of a feature is the computation of the normalized total reduction of the loss function brought by that feature. Following what is done in Scikit-learn, we first need to compute the importance of a node j, n_j, for a single decision tree:

$$n_j = w_j C_j - w_{left,j} C_{left,j} - w_{right,j} C_{right,j},$$

where w_j is the weighted number of samples reaching node j, C_j is the loss function in node j, and left and right are for children nodes from node j. The importance of each feature i, f_i, for the tree is calculated as follows:

$$f_i = \frac{\sum_{j:\ node\ j\ splits\ on\ feature\ i} n_j}{\sum_{k \epsilon all nodes} n_k}.$$

The feature is then normalized and averaged over all the T trees to give the final feature importance, RF_f_i:

$$RF\_f_i = \frac{1}{T} \sum_{t \epsilon all\ trees} \frac{f_{i,t}}{\sum_{j \epsilon all\ features} f_{j,t}}.$$

4.3 Uncertainty quantification

To quantify the uncertainties in addition to the deterministic predictions, two methods were applied and compared: quantile regression random forest and Monte Carlo simulation applied to standard regression random forest. However, our MC-based approach propagates the uncertainty of the input data into the prediction, whereas the QRRF produced probabilistic prediction without considering data uncertainties as input. In both cases, uncertainties are expressed through quantiles: for a continuous distribution function, the α-quantile $Q_\alpha(x)$ is defined such that the probability of Y being smaller than $Q_\alpha(x)$ is exactly equal to α for $X = x$.

$$Q_\alpha(x) = inf\{y:\ P(Y \le y | X = x) \ge \alpha\}. \quad (6)$$

4.3.1 Quantile regression random forests

QRRFs were introduced by Meinshausen (2006) to provide not only the mean, as in the RF, but also the full conditional distribution of the response variable, allowing to construct prediction intervals. Random forest predicts the response variable by a weighted mean over the observations of all leaves, i.e., $E(Y|X = x)$, while quantile regression forest keeps the value of all observations to assess the conditional distribution of Y given by $P(Y \le y|X = x)$. This is achieved by the weighted mean over the observations $1_{\{Y \le y\}}$ using the same weights $w_i(x)$ as those for random forests:

$$P(Y \le y | X = x) = \sum_{i=1}^{n} w_i(x) 1_{\{Y_i \le y\}}. \quad (7)$$

Following Meinshausen (2006), the algorithm for computing Eq. 7 is summarized as follows:

a) k trees were grown as in random forests, but for every leaf of every tree, all observations were retained instead of just their average.
b) For a given $X = x$, x was dropped down all trees. For each observation i, $w_i(x, \theta_t)$ was computed for all trees and so was the average $w_i(x)$.
c) The estimate of the distribution function as in Eq. 7 was computed for all y using weights from step (b).
d) Estimates of the conditional quantiles $Q_\alpha(x)$ were then obtained by plugging the estimates from step (c) in Eq. 6.

Using the RandomForestQuantileRegressor tool from the Scikit-garden Python library, we computed the fifth, 25th, 75th, and 95th quantiles of the QRRF prediction in addition to the mean value.

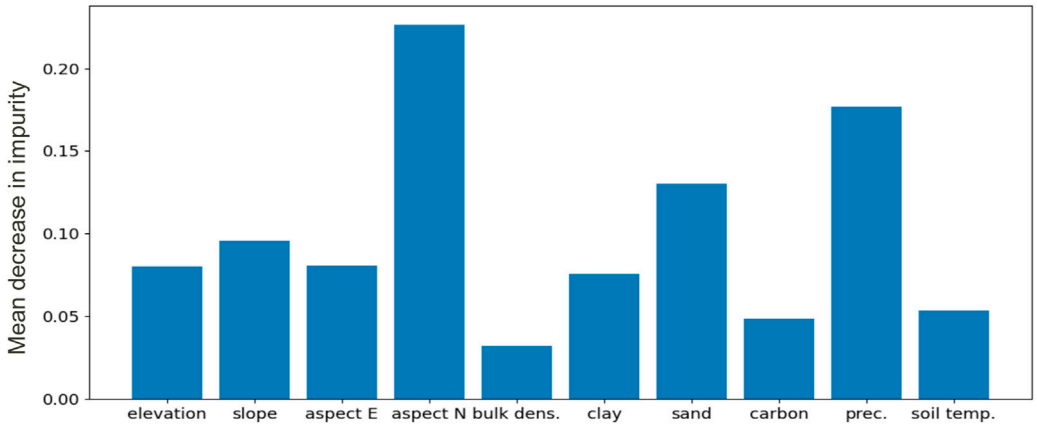

FIGURE 6
Relative feature importance within the training dataset for the deterministic random forest model.

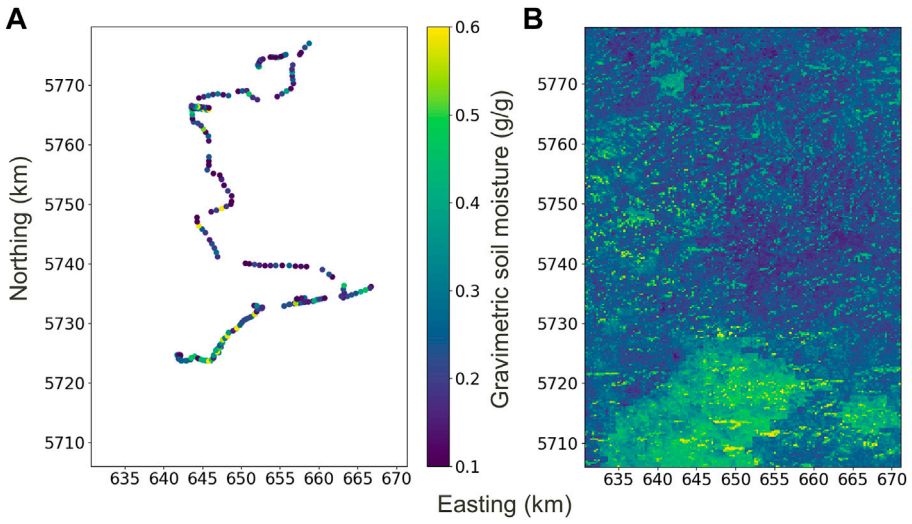

FIGURE 7
Soil moisture values at measurement locations **(A)**, and soil moisture map from random forest prediction of the area of interest **(B)**.

4.3.2 Monte Carlo simulation

Monte Carlo simulation relies on the repeated generation of input random variables from their probability distributions to compute the statistics and estimate the distribution of the output. In this study, we have access to the distribution of the sparse soil moisture (measured dataset of the response variable). Using $m = 1 \ldots 1000$ Monte Carlo steps, we generated, at each step, a random soil moisture training dataset from the probability distributions. Then, we trained a new random forest and performed a deterministic prediction of the dense soil moisture. Thus, we achieved 1,000 RF prediction models $\widehat{\mu_m}(x)$ which were randomly different depending on the input or training data realization drawn from the data probability distributions. Aggregating the results of the 1,000 predictions gave us a probability distribution for all locations of the prediction. From this distribution, statistics such as quantiles could be computed, allowing a direct comparison with the QRRF method.

4.4 Visualization of uncertainty

Quantified uncertainties need to be displayed adequately to be analyzed and interpreted. Visualization of uncertainties for a 1D dataset is easy as values such as variance or quantiles can be represented by curves around the observed or modeled dataset. For a 2D dataset like the spatial data considered here, and in order to avoid dealing with multiple maps, we chose to keep only one parameter to describe the uncertainty, the quartile coefficient of dispersion (CoD). This coefficient is computed using the first and third quartiles (Q_1 and Q_3, respectively):

$$CoD = \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

The CoD, dimensionless, tells us how spread out our data are but is not much affected by outliers since it only depends on the first and third quartiles. Once the CoD is computed, it is superimposed by

10.3389/fenvs.2023.1009191

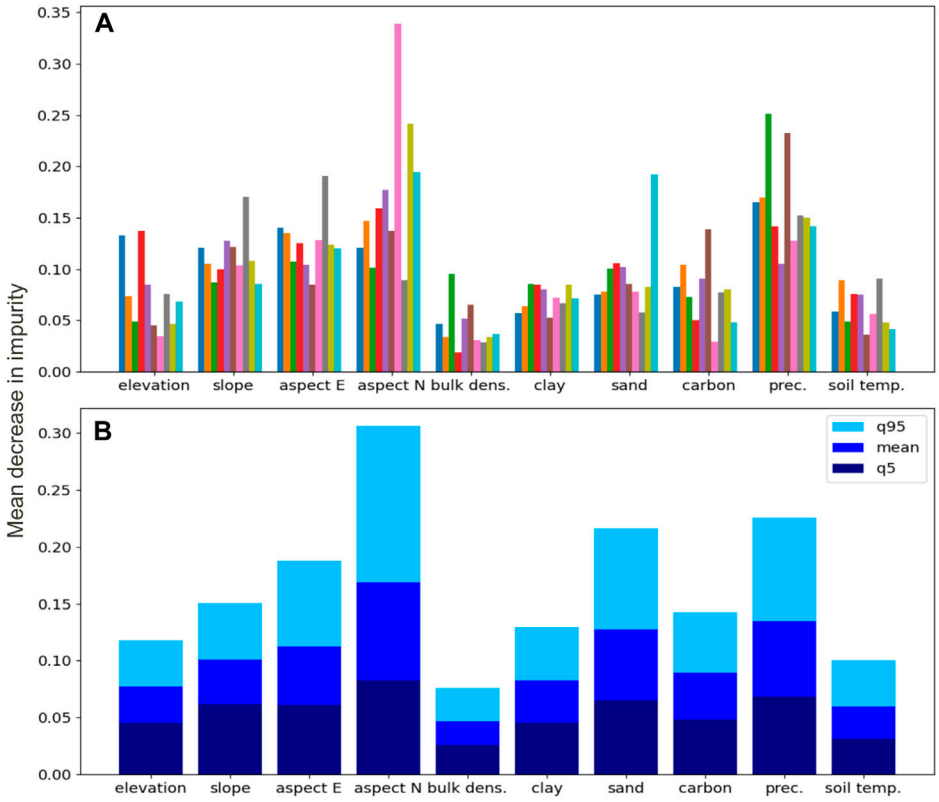

FIGURE 8
Relative feature importance within the training dataset for the first instances of the Monte Carlo approach **(A)** and main statistic values (mean, fifth and 95th percentiles) after 1,000 iterations **(B)**. The large dispersion shows how the feature importance varies with the noise in the input data.

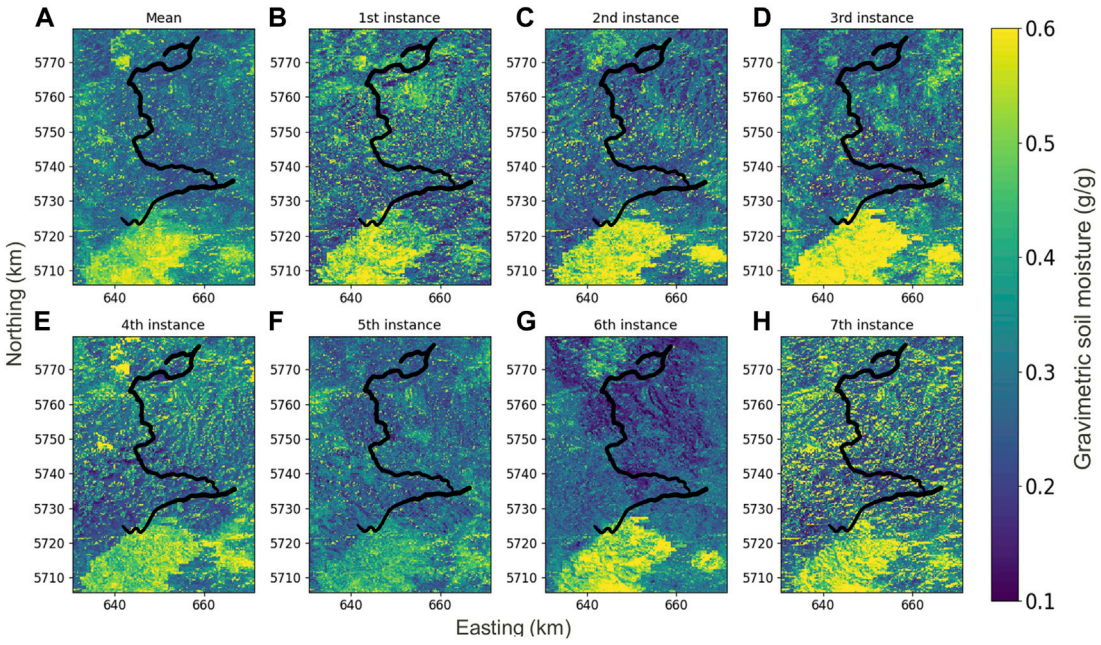

FIGURE 9
Soil moisture maps obtained from random forests through a Monte Carlo approach. The mean soil moisture map after 1,000 iterations **(A)** followed by the first instances of the Monte Carlo simulation **(B–H)** are displayed.

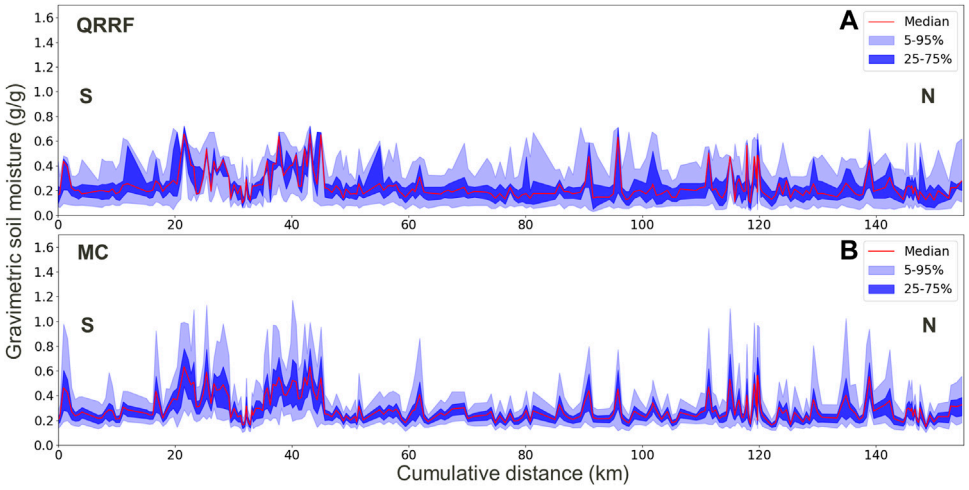

FIGURE 10
Comparison at measurement locations of quantile RF uncertainties **(A)** and input data uncertainties propagated through a Monte Carlo approach **(B)**. Uncertainties are displayed by means of the fifth, 25th, 75th, and 95th percentiles.

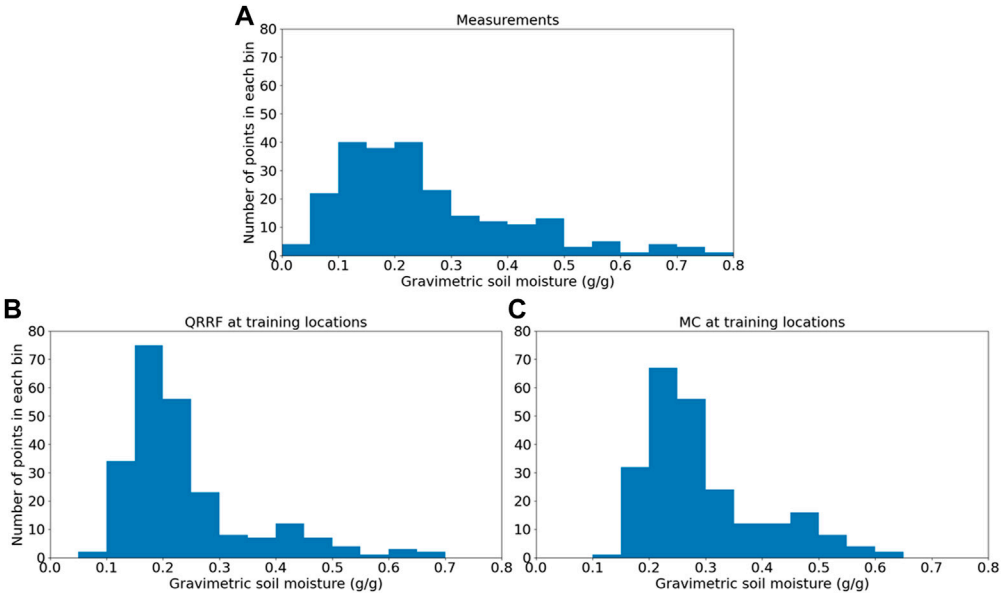

FIGURE 11
Histograms displaying the distribution of the measured soil moisture values **(A)**, soil moisture values at the same locations predicted from QRRF **(B)**, and mean predictions from the MC approach **(C)**.

transparency to the prediction map with a bivariable colormap, where transparent pixels are more uncertain and vivid pixels more certain.

5 Results

A deterministic random forest is first run on the data in order to evaluate the quality of the training dataset. After the training phase, the training performance is checked for convergence. When considering input data uncertainty, we found a variance of 0.01 in computing the MSE of the model depending on the noise contamination of the response variable used.

To check the importance of each of the predictors used for the soil moisture predictions, a feature importance analysis is performed (Figure 6). Overall, three variables have more impact: the slope aspect in the north direction, the precipitation, and the sand proportion of the soil. Although some features seemed to have less impact on the prediction (bulk density, for instance), we kept them all in order to later verify that this feature's importance ranking is stable after the introduction of uncertainty in training data.

The prediction on the area of interest from the rover measurements is shown in Figure 7. Gravimetric soil moisture has been predicted on the grid used for the predictor data, which had a node spacing of 250 m. Some pixels have values quite different from all neighbor pixels, especially high

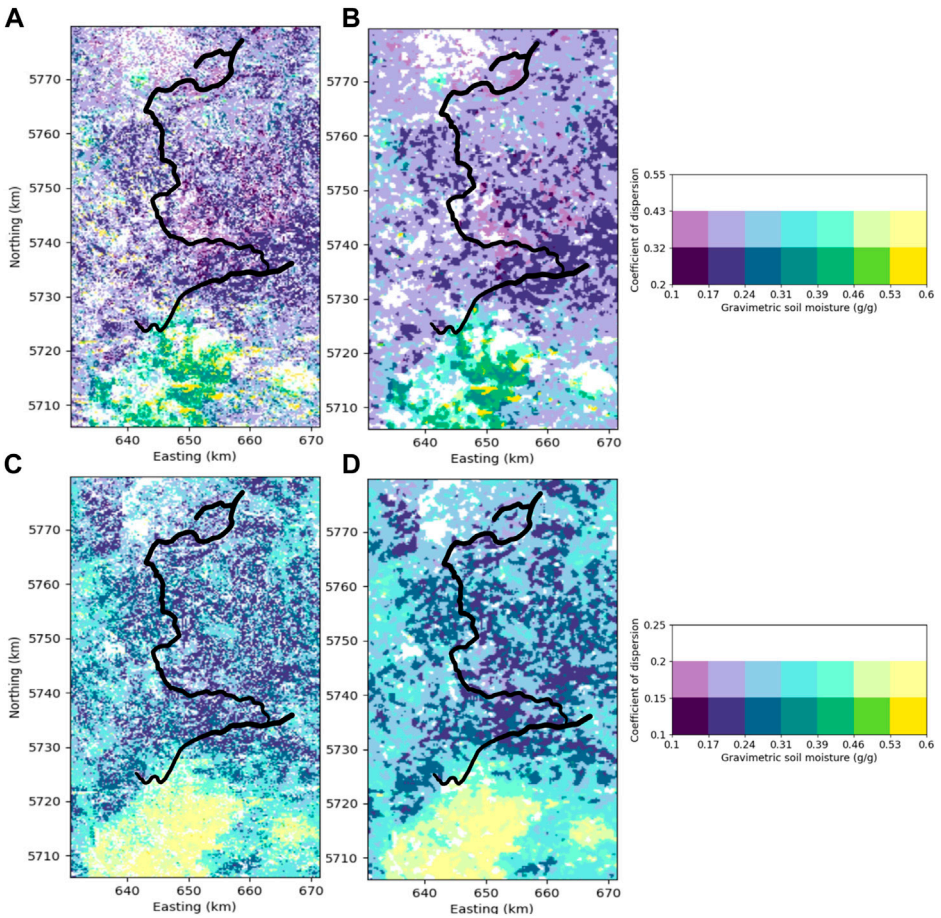

FIGURE 12
Soil moisture map with uncertainty quantification from quantile random forest **(A)**. Transparency is used with the colormap to represent the uncertainty information. **(B)** shows the data from **(A)** after the application of mild spatial smoothing to remove noisy pixels and enhance uncertainty patterns visualization. Soil moisture map with uncertainty quantification from Monte Carlo approach with 1,000 iterations **(C)** and after application of a mild spatial smoothing **(D)**.

isolated values in the south-west of the map, and should be identified as noise, or outlier predictions, since predictors do not have such abrupt changes. We can, however, identify some trends for the soil moisture pattern, specifically the area of high values in the south.

In order to propagate the uncertainties from the sparse training soil moisture values to the full prediction, a Monte Carlo approach is used with 1,000 iterations. For each model, the feature importance analysis is performed to check if it is impacted by the introduction of noise in the training data.

As shown in Figure 8A displaying the 10 first iterations, from one iteration to the other, the feature importance can vary significantly. The seventh iteration, for instance, seems to be largely driven by the slope aspect in the north direction; on the other side, the carbon content of the soil has little effect on this prediction. In most other predictions, however, these two features have a more average impact. This observation gives us the confidence to keep all the predictors for the study, as we just saw that the presence of data noise can change the feature importance ranking. In Figure 8B, we show the statistics (mean, fifth and 95th quantiles) of the feature importance after the 1,000 iterations. These statistics confirmed the high variability of the feature importance and stayed coherent with the first iterations from (A). Bulk density always has lower statistics than the other features; however, since its importance can still be higher than those of other features in some predictions within our Monte Carlo approach, it cannot be completely regarded as a predictor. The slope aspect to the north on the opposite always has a significant impact and can even have a very high impact in some cases. For the features with more impact, slope aspect in the north direction, and sand and precipitations, we observe a great gap between the fifth and the 95th quantiles.

The introduction of noise also generates substantial variability in the prediction maps (Figure 9). In comparison to the mean prediction generated by averaging the 1,000 predictions (panel A), some individual predictions, such as the fifth prediction (panel E), have much less high soil moisture values (i.e., above 0.6), especially in the southern and the north-western parts. On the opposite side, other predictions have much more higher values (seventh prediction, panel H, for instance, in Figure 9), where we observe sparse high values everywhere and, to a large extent, in the middle and eastern parts, where the mean prediction has only very isolated high values in these areas. This high variability is due to the high uncertainty of some of the input data, which is up to 40%.

From all these predictions, the mean and the fifth, 25th, 75th, and 95th quantiles were computed. With this information, we can compare

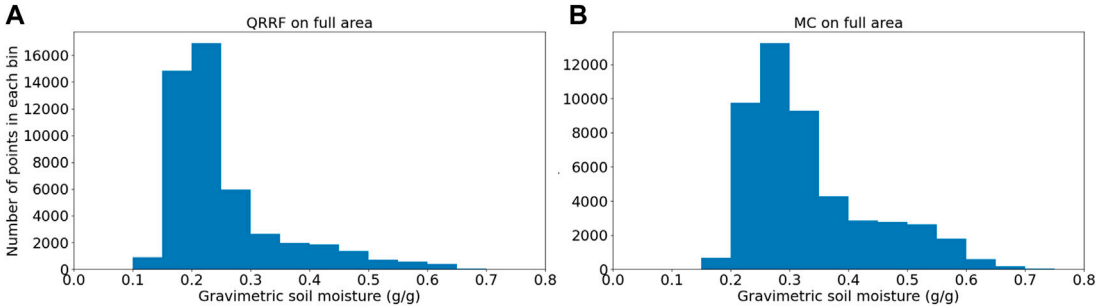

FIGURE 13
Histograms displaying the distribution of the predicted soil moisture values on the full area for the QRRF prediction **(A)** and the MC after 1,000 iterations **(B)**.

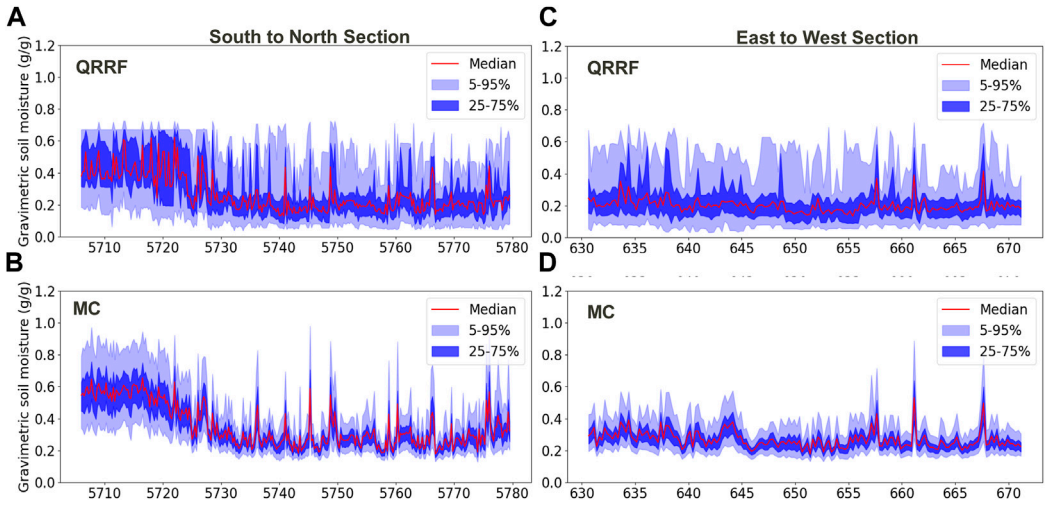

FIGURE 14
Soil moisture mean prediction and fifth, 25th, 75th, and 95th percentiles along the centered south/north section [**(A)** QRRF and **(B)** MC] and the centered west/east section [**(C)** QRRF and **(D)** MC].

the statistics from the Monte Carlo approach to the quantiles obtained from a quantile RF. This is shown in Figure 10, where the mean and the quantiles are shown for the predictions achieved by both methods at the training locations to be comparable with the input data and original uncertainties also (Figure 3).

For the QRRF prediction (A in Figure 10), we see that the predictions never exceed 0.8, which is the highest soil moisture value found in the original training dataset. The random forest is indeed unable to extrapolate values outside the training range, and that is why uncertainties leading to higher soil moisture values are not well recovered. Due to this limitation, for the highest soil moisture values, the 75th and 95th quantiles are almost equal to the mean value. As for the MC approach (B in Figure 10), a wider range of values can be predicted since more extreme values are seen during the multiple training phases. Compared to Figure 3, we see that MC quantiles are relatively similar to input data uncertainties, and the global trend is recovered with a good matching between more certain and more uncertain areas. Although QRRF predicts more homogeneous quantiles over the training dataset, areas with higher uncertainties (for instance, between 20 and 45 km or 110–120 km) are nonetheless recovered to a certain extent. This is different for the areas with lower uncertainties which are not much recovered in the QRRF output.

Histograms in Figure 11 show the distribution of the soil moisture values at the measurement locations for the measured values (A) and the predicted values from QRRF (B) and the MC approach (C). For the MC approach, to be comparable with the other histograms, the median value from all simulations is computed and used for the histogram.

The quantile prediction is then used on the full study area once again for both methods. The results are plotted on a map with a bivariable colormap, with the color representing the gravimetric soil moisture and transparency of the coefficient of dispersion (Figure 12). This way, vivid colors represent a higher certainty for the computed soil moisture, whereas dull, pale colors stand for uncertain predictions. As the prediction result are not very trustworthy when uncertainty is very high, the transparency is used indifferently for all uncertain predictions.

The prediction from the MC approach (1,000 repetitions) is shown in Figures 12C,D, while the one from QRRF is shown in

Figures 12A,B. Similar to the observations on the profiles (Figure 10), QRRF has a wider range of uncertainties (the coefficient of dispersion for QRRF is almost twice as big as that for MC) and fewer contrasts compared to MC. The mean values are also lower for QRRF due to the training with less extreme values, as explained previously. With this full prediction, the areas with higher uncertainties are again predicted in similar locations in both methods (west, north-west, maybe south and south-east also). This highlights the ability of both methods to identify very uncertain values. For the areas with lower uncertainties, however, the QRRF is usually not able to identify them very well and provides very homogeneous uncertainties. The direct propagation of the input uncertainty through Monte Carlo performs better in this task for very small to very large prediction intervals.

Another comparison between both methods is performed through histograms to analyze the distribution of predicted values on the full area (Figure 13). We can see that QRRF predictions have a prominent concentration of predicted values around 0.2–0.25 g/g, whereas MC predictions have more distributed values over a wider range of soil moisture values. As already seen in Figure 11, MC predictions have fewer extreme values as we only considered the median values for the comparison.

Two profiles are extracted from the full prediction, one centered section from east to west and another centered section from south to north (Figure 14). These profiles allow us to better see the heterogeneity of MC uncertainty, especially in the south/north section (A and B), whereas QRRF uncertainty are wide roughly everywhere, MC uncertainty are wide in the southern part and are smaller otherwise. On the east/west section (C and D), we observe some points with very high uncertainty for the 75th percentile and overall much more extremely high prediction (highlighted with the 95th percentile) in the QRRF prediction. MC predictions on this section, however, are more homogenous along the section.

6 Discussion and conclusion

In this paper, we show a successful application of the Monte Carlo approach to propagate response variable uncertainties through the solution of a multiple regression problem into predicted maps of gravimetric soil moisture. The process allows us to predict soil moisture on a dense spatial grid with quantification of prediction uncertainty rooted in response variable uncertainty. Despite ignoring other sources of uncertainty in the MRP, this helps us to spot places where prediction is more trustworthy or, on the contrary, less reliable. The found uncertainties represent a lower limit.

Consideration of further sources of uncertainty, e.g., predictor data uncertainty or uncertainties related to methodological choices of how to compute the regression models might overlap with the uncertainties computed here. Our approach relies on the quantified uncertainties of the training response variable. In contrast, QRRF considers uncertainty representing some aspects of model uncertainty. MC gives a more spatially contrasted prediction than the QRRF and partly exceeds the uncertainty of QRRF. While the QRRF prediction results reveal a first-order correlation of uncertainty with gravimetric soil moisture, e.g., high soil moisture values that go along with high uncertainties; the relation between uncertainty and soil moisture is more complex in the MC approach. Particularly, for low soil moisture readings, the MC approach propagating data uncertainty into the prediction is more suitable than the QRRF approach to assess the prediction uncertainty not overoptimistically, to support proper information extraction from the predictions and avoid over-interpretation.

The QRRF approach is not well suited to clearly identify areas of lower or higher uncertainties since input data taken up in the training ignore uncertainty. It illustrates a different aspect of uncertainty than the Monte Carlo approach and is more related to the ability of the model to learn a pattern for the target variable prediction with the provided training data. Thus, we judge its provided uncertainty to be merely related to internal functionality and chosen settings of the random forest algorithm used to solve the MRP. Accordingly, the quantified uncertainty will be more spatially homogeneous. Since QRRF is a method routinely used to quantify the spatial uncertainty of soil maps of various essential state variables, a coupling of the MC approach with the QRRF might have the potential to reach a more reliable uncertainty assessment. However, the study of model uncertainty and the general suitability of QRRF for model uncertainty quantification was not the goal of this study. When analyzing the results presented in this paper, we have to keep in mind that the model performance is not very good due to the small size of the dataset, and it would be interesting to perform a similar comparison between the MC approach and the QRRF with a bigger training dataset. However, the small volume of training data makes the MC approach computationally tractable, which may be a limitation of the MC approach with much bigger datasets.

A further step in the future to improve this analysis for the Monte Carlo approach is to also include predictor uncertainties for a more comprehensive uncertainty quantification of the soil map product. However, since the number of predictors is about 10, a simple MC approach as we used here might suffer the curse of dimensionality, and the sampling step will, therefore, require more attention. Furthermore, the study shows the importance of technological development for reducing the uncertainty in training data.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article and on request from the author if they are not subject to third party restrictions.

Author contributions

SD and HP developed the concepts of the presented idea. MS provided the data and support with all CRNS-related questions. SD took the lead in writing the manuscript. PD supervised the project. All authors provided critical feedback and helped in writing the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Adab, H., Morbidelli, R., Saltalippi, C., Moradian, M., and Ghalhari, G. A. F. (2020). Machine learning to estimate surface soil moisture from remote sensing data. *Water* 12, 3223. doi:10.3390/w12113223

Baake, K. (2018). Quantifying uncertainty of random forest predictions: A digital soil mapping case study, Thesis report GIRS-2017-14. Wageningen, Netherlands: Wageningen University .

Baroni, G., Scheiffele, L. M., Schrön, M., Ingwersen, J., and Oswald, S. E. (2018). Uncertainty, sensitivity and improvements in soil moisture estimation with cosmic-ray neutron sensing. *Journal of Hydrology* 564, 873–887. doi:10.1016/j.jhydrol.2018.07.053

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Belmont, NY, USA: Wadsworth.

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

Carranza, C. D., Nolet, C., Pezij, M., and van der Ploeg, M. (2021). Root zone soil moisture estimation with Random Forest. *Journal of Hydrology* 593, 125840. doi:10.1016/j.jhydrol.2020.125840

Desilets, D., Zreda, M., and Ferré, T. P. A. (2010). Nature's neutron probe: Land surface hydrology at an elusive scale with cosmic rays. *Water Resour. Res.* 46, W11505. doi:10.1029/2009WR008726

Durbin, J., and Koopman, S. J. (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika* 84, 669–684. doi:10.1093/biomet/84.3.669

Gruber, A., and Peng, J. (2022). "Remote sensing of soil moisture," in *Reference module in Earth systems and environmental sciences* (Amsterdam, Netherlands: Elsevier). doi:10.1016/B978-0-12-822974-3.00019-7

Hawdon, A., McJannet, D., and Wallace, J. (2014). Calibration and correction procedures for cosmic-ray neutron soil moisture probes located across Australia. *Water Resources Research* 50, 5029–5043. doi:10.1002/2013WR015138

Hengl, T., Heuvelink, G. B. M., and Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma* 120, 75–93. doi:10.1016/j.geoderma.2003.08.018

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

Heuvelink, G. B. M. (1998). *Error propagation in environmental Modelling with GIS*. London, UK: CRC Press. doi:10.4324/9780203016114

Heuvelink, G. B. M. (2014). "Uncertainty quantification of GlobalSoilMap products," in *GlobalSoilMap* (CRC Press), 335–340. doi:10.1201/b16500-62

Heuvelink, G., and Webster, R. (2001). Modelling soil variation: Past, present and future. *Geoderma* 100, 269–301. doi:10.1016/S0016-7061(01)00025-8

Howarth, R. J. (2001). A history of regression and related model-fitting in the Earth sciences (1636?-2000). *Natural Resources Research* 10, 241–286. doi:10.1023/A:1013928826796

Jakobi, J., Huisman, J. A., Schrön, M., Fiedler, J., Brogi, C., Vereecken, H., et al. (2020). Error estimation for soil moisture measurements with cosmic ray neutron sensing and implications for rover surveys. *Front. Water* 2, 10. doi:10.3389/frwa.2020.00010

JCGM (2008). Guide to the expression of uncertainty in measurement. http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf.

Kasner, M. (2016). Kalibrierung mobiler Cosmic-Ray-Neutronen-Messung in Bezugauf Vegetation und Bodeneigenschaften zur Abschätzung räumlicher Boden-feuchte. MSc thesis, Halle (Saale), Germany: Martin-Luther-Universität Halle-Wittenberg .

Köhli, M., Weimar, J., Schrön, M., Schmidt, U., and Schmidt, U. (2021). Soil moisture and air humidity dependence of the above-ground cosmic-ray neutron intensity. *Frontiers in Water* 2, 66. doi:10.3389/frwa.2020.544847

Korres, W., Koyama, C., Fiener, P., and Schneider, K. (2010). Analysis of surface soil moisture patterns in agricultural landscapes using Empirical Orthogonal Functions. *Hydrology and Earth System Sciences* 14. doi:10.5194/hessd-6-5565-2009

Kuhn, M., and Johnson, K. (2013). *Applied predictive modeling*. New York, NY, USA: Springer.

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., Martin, M. P., and Saby, N. P. (2019). How far can the uncertainty on a digital soil map be known?: A numerical experiment using pseudo values of clay content obtained from vis-SWIR hyperspectral imagery. *Geoderma* 337, 1320–1328. doi:10.1016/j.geoderma.2018.08.024

Lagacherie, P. (2008). Digital soil mapping: A state of the art, *Digital soil Mapping With Limited Data*, 181. Dordrecht, Netherland: Springer-Verlag, 978–981.

Lakshmi, V., Jackson, T. J., and Zehrfuhs, D. (2003). Soil moisture-temperature relationships: Results from two field experiments. *Hydrological Processes* 17, 3041–3057. doi:10.1002/hyp.1275

Lorenzetti, R., Roberto, B., Fantappiè, M., L'Abate, G., and Costantini, E. (2015). Comparing data mining and deterministic pedology to assess the frequency of WRB reference soil groups in the legend of small scale maps. *Geoderma* 237–238, 237–245. doi:10.1016/j.geoderma.2014.09.006

McBratney, A. B., Mendonça Santos, M. L., and Minasny, B. (2003). On digital soil mapping. *Geoderma* 117, 3–52. doi:10.1016/S0016-7061(03)00223-4

McJannet, D., Hawdon, A., Baker, B., Renzullo, L., and Searle, R. (2017). Multiscale soil moisture estimates using static and roving cosmic-ray soil moisture sensors. *Hydrology and Earth System Sciences* 21, 6049–6067. doi:10.5194/hess-21-6049-2017

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research* 7, 983–999.

Mentch, L., and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research* 17, 841–881. doi:10.48550/arxiv.1404.6473

Nauman, T. W., and Duniway, M. C. (2019). Relative prediction intervals reveal larger uncertainty in 3D approaches to predictive digital soil mapping of soil properties with legacy data. *Geoderma* 347, 170–184. doi:10.1016/j.geoderma.2019.03.037

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., et al. (2018). Evaluation of digital soil mapping approaches with large sets of environmental covariates. *Soil* 4, 1–22. doi:10.5194/soil-4-1-2018

Paasche, H., Gross, M., Lüttgau, J., Greenberg, D. S., and Weigel, T. (2021). To the brave scientists: Aren't we strong enough to stand (and profit from) uncertainty in Earth system measurement and modelling? *Geoscience Data Journal* 9, 393–399. doi:10.1002/gdj3.132

Pérez-Díaz, L., Alcalde, J., and Bond, C. (2020). Introduction: Handling uncertainty in the geosciences: Identification, mitigation and communication. *Solid Earth* 11, 889–897. doi:10.5194/se-11-889-2020

Poggio, L., de Sousa, L. M., Batjes, N. L., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., et al. (2021). SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *Soil* 7, 217–240. doi:10.5194/soil-7-217-2021

Schmidt, A., Dabas, M., and Apostolos, S. (2020). Dreaming of perfect data: Characterizing noise in archaeo-geophysical measurements. *Geosciences* 10, 382. doi:10.3390/geosciences10100382

Schrön, M., Oswald, S. E., Zacharias, S., Kasner, M., Dietrich, P., and Attinger, S. (2021). Neutrons on rails: Transregional monitoring of soil moisture and snow water equivalent. *Geophysical Research Letters* 48, 24. doi:10.1029/2021GL093924

Schrön, M., Rosolem, R., Köhli, M., Piussi, L., Schröter, I., Iwema, J., et al. (2018). Cosmic-ray neutron rover surveys of field soil moisture and the influence of roads. *Water Resources Research* 54, 6441–6459. doi:10.1029/2017WR021719

Schrön, M., Zacharias, S., Köhli, M., Weimar, J., and Dietrich, P. (2015). Monitoring environmental water with ground albedo neutrons from cosmic rays. The 34th International Cosmic Ray Conference Hague, Netherlands 236, 231. doi:10.22323/1.236.0231

Schröter, I., Paasche, H., Dietrich, P., and Wollschläger, U. (2015). Estimation of catchment-scale soil moisture patterns based on terrain data and sparse TDR measurements using a fuzzy C-means clustering approach. *Vadose Zone Journal* 14. doi:10.2136/vzj2015.01.0008

Sinha, S., Rode, M., and Borchardt, D. (2016). Examining runoff generation processes in the Selke catchment in central Germany: Insights from data and semi-distributed numerical model. *Journal of Hydrology Regional Studies* 7, 38–54. doi:10.1016/j.ejrh.2016.06.002

van der Westhuizen, S., Heuvelink, G. B. M., Hofmeyr, D. P., and Poggio, L. (2022). Measurement error-filtered machine learning in digital soil mapping. *Spatial Statistics* 47, 100572. doi:10.1016/j.spasta.2021.100572

Vaysse, K., and Lagacherie, P. (2017). Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma* 291, 55–64. doi:10.1016/j.geoderma.2016.12.017

Wadoux, A. M. J.-C., Padarian, J., and Minasny, B. (2019). Multi-source data integration for soil mapping using deep learning. *Soil* 5, 107–119. doi:10.5194/soil-5-107-2019

Wager, S., Hastie, T., and Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research* 15, 1625–1651.

Western, A. W., Grayson, R. B., Blöschl, G., Willgoose, G. R., and Mc-Mahon, T. A. (1999a). Observed spatial organization of soil moisture and its relation to terrain indices. *Water Resources Research* 35, 797–810. doi:10.1029/1998WR900065

Western, A. W., Zhou, S.-L., Grayson, R. B., McMahon, T. A., Blöschl, G., and Wilson, D. J. (2004). Spatial correlation of soil moisture in small catchments and its relationship to dominant spatial hydrological processes. *J. Hydrol.* 286, 113–134. doi:10.1016/j.jhydrol.2003.09.014

Winter, C., Lutz, S. R., Musolff, A., Kumar, R., Weber, M., and Fleckenstein, J. H. (2021). Disentangling the impact of catchment heterogeneity on nitrate export dynamics from event to long-term time scales. *Water Resources Research* 57, e2020WR027992. doi:10.1029/2020WR027992

Wollschläger, U., Attinger, S., Borchardt, D., Brauns, M., Cuntz, M., Dietrich, P., et al. (2016). The bode hydrological observatory: A platform for integrated, interdisciplinary hydro-ecological research within the TERENO harz/central German Lowland observatory. *Environmental Earth Sciences* 76, 29. doi:10.1007/s12665-016-6327-5

Yang, X., Jomaa, S., and Rode, M. (2019). Sensitivity analysis of fully distributed parameterization reveals insights into heterogeneous catchment responses for water quality modeling. *Water Resources Research* 55, 10935–10953. doi:10.1029/2019WR025575

Yang, X., Jomaa, S., Zink, M., Fleckenstein, J. H., Borchardt, D., and Rode, M. (2018). A new fully distributed model of nitrate transport and removal at catchment scale. *Water Resources Research* 54, 5856–5877. doi:10.1029/2017WR022380

Zacharias, S., Bogena, H., Samaniego, L., Mauder, M., Fuß, R., Pütz, T., et al. (2011). A network of terrestrial environmental observatories in Germany. *Vadose Zone Journal* 10, 955–973. doi:10.2136/vzj2010.0139

Zhang, H., Nettleton, D., and Zhu, Z. (2017). Regression-enhanced random forests. *JSM proceedings, section on statistical learning and data science*. Alexandria, VA, USA: American Statistical Association, 636647.

Zreda, M., Shuttleworth, W. J., Zeng, X., Zweck, C., Desilets, D., Franz, T. E., et al. (2012). Cosmos: The cosmic-ray soil moisture observing system. *Hydrology and Earth System Sciences* 16, 4079–4099. doi:10.5194/hess-16-4079-2012