Geoderma 437 (2023) 116585

Contents lists available at ScienceDirect

Geoderma

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

Validation of uncertainty predictions in digital soil mapping

Jonas Schmidinger[a,b,*], Gerard B.M. Heuvelink[a,c]

[a] Wageningen University and Research, Soil Geography and Landscape Group, Wageningen, the Netherlands
[b] Leibniz Institute for Agricultural Engineering and Bioeconomy e.V. (ATB), Department of Agromechatronics, Potsdam, Germany
[c] ISRIC-World Soil Information, Wageningen, the Netherlands

ARTICLE INFO

Handling Editor: Budiman Minasny

Keywords:
Validation
Digital soil mapping
Uncertainty
Machine learning
Proper scoring rules
Quantile regression

ABSTRACT

It is quite common in digital soil mapping (DSM) to quantify the uncertainty of issued predictions, that is to make probabilistic predictions. Yet, little attention has been paid to its validation. Probabilistic predictions are only of value for end users if they are reliable and ideally also sharp. Reliability refers to the consistency between predicted conditional probabilities and observed frequencies of independent test data. Sharpness refers to the concentration of a conditional probability distribution function, i.e. its narrowness. The prediction interval coverage probability (PICP) is currently used in DSM to validate the reliability of prediction intervals but it is ignorant of a potential one-sided bias of its boundaries. Therefore, we propose to extend the current validation procedure with metrics used in the broader probabilistic literature. These metrics not only evaluate probabilistic predictions in prediction interval format but also quantiles or full conditional probability distributions. We suggest the quantile coverage probability (QCP) and probability integral transform (PIT) histogram as alternatives to PICP and proper scoring rules for relative comparisons of competing probabilistic models. As scoring rules, we present the interval score (IS) and the continuous ranked probability score (CRPS), which can be decomposed into a reliability part (RELI). We illustrated the use of these metrics in a case study using soil pH and soil organic carbon from the LUCAS-soil database. Thereby, probabilistic predictions of five different models were compared: a reference null model (NM), quantile regression forest (QRF), quantile regression post-processing of a random forest (QRPP RF), kriging with external drift (KED) and quantile regression neural network (QRNN). For KED and QRNN, one-sided bias was found. This was not apparent from PICP but was shown by use of the PIT histogram and QCP. RELI summarized the trends found in QCP, PICP and PIT histograms to one numerical value. CRPS and IS were especially harsh to outliers and low sharpness. According to CRPS and IS, the best probabilistic predictions were obtained by QRF and QRPP RF and the worst by NM.

1. Introduction

Soils are of great importance to humankind since they provide various ecosystem services that contribute to food production, climate mitigation and air- and water quality (Keesstra et al., 2016). In order to maintain these services, soil as a resource has to be adequately managed and protected. This requires quantitative soil information at high spatial resolution, prompting the increasing popularity of digital soil mapping (DSM) (Chen et al., 2022). DSM creates soil maps through statistical inferences from a prediction model, using exhaustively accessible environmental covariates as predictors and soil sample data for model training (McBratney et al., 2003). Unavoidably, these predictions and thus the generated soil maps are not error-free. Map error originates from a variety of sources but most importantly it comes from the inability of the covariates to explain all soil spatial variation (Nelson et al., 2011). Other sources of error include the limited ability of a model to exploit all information provided by the covariates, a too-small training sample size, and measurement errors in the training data.

Estimation of the overall error can be done using a design-based approach, with independent test data obtained from probability sampling (Brus et al., 2011). Using this approach, map predictions are compared to the independent observations. The model performance, i.e. map accuracy, can then be quantified by well-established validation metrics such as the mean error (ME), root-mean-squared error (RMSE), and Nash-Sutcliffe model efficiency coefficient (MEC) (Piikki et al., 2021).

End users might not only be interested in the overall map accuracy but might require information about the accuracy at each and every

[*] Corresponding author at: Soil Geography and Landscape Group, Wageningen University, Droevendaalsesteeg 3, 6708BP Wageningen, the Netherlands.
E-mail addresses: Jonas.Schmidinger@outlook.com (J. Schmidinger), gerard.heuvelink@wur.nl (G.B.M. Heuvelink).

https://doi.org/10.1016/j.geoderma.2023.116585
Received 8 April 2023; Received in revised form 22 June 2023; Accepted 27 June 2023
Available online 11 July 2023

location in the mapped study area. In such case, a design-based statistical inference is not suitable because it only provides summary measures of the map accuracy. However, with a model-based approach (Heuvelink, 2018) location-specific information about the prediction accuracy can be derived through the use of a probabilistic prediction model. A probabilistic prediction model goes beyond point prediction and estimates the entire conditional probability distribution of a soil property of interest, either directly or from a large set of conditional quantiles (Lauret et al., 2019). We refer to them as predictive distributions. They are generated for every location in the area of interest, in which the mean of the predictive distribution is typically used as a point prediction. The predictive distribution defines the probability of obtaining a large or small prediction error. A narrow, also called sharp, predictive distribution indicates that the point prediction is likely close to the true value. In such case, we are confident about the obtained point prediction and do not expect to have a large prediction error. With a wider predictive distribution, it cannot be ruled out that the prediction error is large, meaning that we are more uncertain if the true value is close to the predicted value. In DSM, we usually refer to this general concept as uncertainty (Heuvelink, 2018). In the following, we will use the more general term ‘probabilistic prediction’ as a synonym for ‘uncertainty prediction’.

While uncertainty is completely characterized by a predictive distribution, often it is summarized through a prediction interval (PI) for a more intuitive and practical interpretation. PI indicates a range in which the true value is expected to be found, given an assigned probability. Usually, a 90% prediction interval (PI) is used in DSM (Chen et al., 2022). For instance, if the 90% PI of the pH of the soil at some location is given by [5.3, 7.1], we claim that there is a chance of 90% that the true soil pH is between the lower limit of 5.3 and the upper limit of 7.1. Note that it is important to clearly convey the meaning of prediction intervals to end users. For instance, some might wrongfully conclude from it that the pH is uniformly distributed between the lower and upper limit.

Indicating the uncertainty, usually in the form of a PI, has multiple advantages, such as: (i) preventing end users from getting a wrong sense about the accuracy of a soil map, thus allowing them to decide if the quality of the map is sufficient for the intended purpose (Heuvelink, 2018); (ii) allowing uncertainty propagation if the soil map is further used as input in other simulations or models (Heuvelink, 1998); and (iii) enabling to consider uncertainty in decision making (Breure et al., 2022; Lark et al., 2022). Because of these reasons, it is strongly encouraged to deliver the underlying uncertainty next to the actual predicted soil attributes.

Traditionally, much attention has been paid in DSM to quantify uncertainty by kriging models (Goovaerts, 2001), such as ordinary kriging or kriging with external drift (Webster and Oliver, 2007), or use of the empirical model (Malone et al., 2011). However, machine learning algorithms that are able to predict conditional quantiles are getting increasingly popular in DSM (Kasraei et al., 2021; Lagacherie et al., 2019; Vaysse and Lagacherie, 2017). Two examples of such techniques are quantile regression forest (Meinshausen, 2006) and quantile regression neural network (Cannon, 2011), which are probabilistic adaptions of a random forest and an artificial neural network, respectively. Recently, Kasraei et al. (2021) introduced quantile regression post-processing, which makes use of a quantile regression (Koenker and Hallock, 2001) implemented in a two-step algorithm.

Even though often disregarded, probabilistic predictions must be validated too since a poor uncertainty map could encourage harmful decisions if used in practice. The validation of probabilistic predictions is more complicated than the validation of point predictions as the former are characterized by probabilities and can occur in different forms such as PIs, quantiles, predictive distribution functions, or a mixture of them. Two general attributes that are usually evaluated for probabilistic predictions are reliability and sharpness (Gneiting and Raftery, 2007). Reliability, also known as probabilistic calibration, is a measure of the consistency between the predicted probability and the empirical frequency of independent test data. A probabilistic prediction should also be informative, which can be expressed by its sharpness. Sharpness refers to the concentration of a predictive distribution. Hence, high sharpness is characterized by a narrow PI or predictive distribution. In Gneiting and Raftery (2007) the goal for a probabilistic prediction is defined as to “maximize the sharpness of the predictive distributions subject to reliability”. In DSM we usually measure sharpness by the prediction interval width and validate reliability with the prediction interval coverage probability (PICP) (Goovaerts, 2001; Malone et al., 2011). PICP evaluates if the probability assigned to the PIs is equal to the frequency of empirical test data within the PIs. Various studies compared the reliability of probabilistic prediction models frequently used in DSM based on PICP (e.g. Kasraei et al., 2021; Szatmári and Pásztor, 2019; Vaysse and Lagacherie, 2017). Vaysse and Lagacherie (2017) and Szatmári and Pásztor (2019) reported poor reliability due to suboptimal PICP values with kriging and it was outperformed by quantile regression forest. Contrarily, Kasraei et al. (2021) obtained inconsistent PICP values for quantile regression forest depending on the predicted soil property, whereas quantile regression post-processing combined to various machine learning models had a more stable performance.

One reason for these ambiguous results may lie in the use of PICP as a validation metric. Pinson and Tastu (2014) pointed out that PICP is not an optimal metric to measure reliability, since it cannot account for one-sided bias. One-sided bias refers to a case where two quantile predictions that define a PI are simultaneously shifted either negatively or positively, as explained in more detail in Section 2.6.2. Therefore, it is of interest to expand the current validation procedure in DSM with more validation metrics that do not have this disadvantage. In other academic fields, in which probabilistic predictions have a longer tradition, its validation is more comprehensive (Bracher et al., 2021; Lauret et al., 2019; Pinson et al., 2007; Zhang et al., 2014). These fields naturally include a broader set of validation metrics such as proper scoring rules (Gneiting and Raftery, 2007), probability integral transform histograms and the quantile coverage probability (Lauret et al., 2019).

The overall objective of this study is to introduce well-established concepts for the validation of probabilistic predictions from other academic fields to DSM. In a case study, using Land Use and Coverage Area Frame Survey (LUCAS) data, their added value will be illustrated for several probabilistic prediction models relevant to DSM. The performance of these models will then be compared on the basis of the old- and the newly introduced metrics.

2. Materials and methods

2.1. Study area, soil data and covariates

We used the soil pH and soil organic carbon (SOC) data from LUCAS-soil 2015 in the area of Germany and Benelux, which consisted of 2,018 data points. LUCAS-soil contains various soil attributes for all countries of the European Union. The large size, open license and consistent methodology makes LUCAS-soil attractive for testing new methods in DSM. Each soil sample of LUCAS-soil is a composite sample of five topsoil subsamples (0–20 cm) (Orgiazzi et al., 2018). pH was measured according to ISO 10390:1994 (ISO, 1994) with a glass electrode both in water and in a calcium-chloride solution. In this study, we only used pH measured in calcium-chloride. SOC was obtained according to ISO 10694:1995 (ISO, 1995) through the determination of the sample weight loss after dry combustion and removal of carbonates. Further, we log-transformed SOC to log(SOC) to remove skewness. The study area and sampling locations and values of pH can be found in Fig. 1. Those of log(SOC) are given in Fig. A1 in the Supplementary Information (SI).

We used the pre-processed covariates from Poggio et al. (2021), which consisted of various environmental factors important in the context of soil formation. These were, among others, vegetation indices, climate variables, land cover, terrain morphology and lithology.

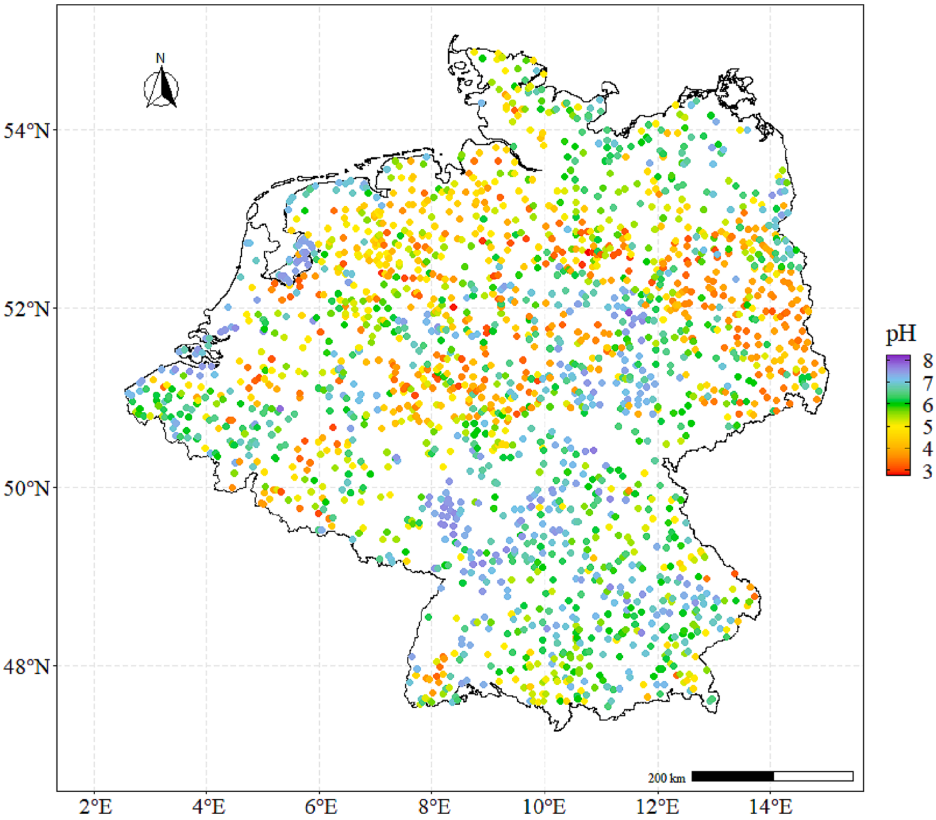

Fig. 1. LUCAS-soil sampling sites in Germany and Benelux with color-coded values of soil pH.

Additionally, we addressed intercorrelation by randomly dropping one covariate of each covariate pair that had a Pearson correlation bigger than 0.9 and eliminated covariates with near-zero variance. This led to a list of 89 covariates. We excluded two soil samples in areas which had missing values for some covariates, leading to a total of $N = 2016$ soil samples used in this study.

2.2. *Study design*

In an outer loop, the original dataset of size N was fully randomly split for S times ($S = 25$) into a training set (75%) for model fitting and a test set (25%) for the validation of model performance. Five prediction models with probabilistic capabilities were trained (Section 2.3) to issue both point predictions and probabilistic predictions (Section 2.4) for pH and log(SOC) at every sample site of the test set. Prior to model training, for some of these models, a hyperparameter selection or a step-wise variable selection grounded on the Akaike Information Criterion was implemented. The hyperparameter selection was based on a grid search within an inner loop with K-fold cross-validation ($K = 5$) of the training set. Note, that the selection was executed on the basis of optimizing point prediction performances, not probabilistic prediction performances. After the training, the test set was used to validate the point predictions with standard validation metrics (Section 2.5) and the probabilistic predictions with the PICP and newly proposed validation metrics (Section 2.6). Finally, the performances obtained from the outer loop were stored and aggregated over the S repetitions. The whole study design is conceptualized in Fig. 2.

2.3. *Probabilistic prediction models*

2.3.1. *Null model*

The null model (NM) uses the mean of the training set for the point predictions and the empirical cumulative distribution function (CDF) of the training set as a predictive CDF for the probabilistic predictions. Note that this implies that both the point predictions and the probabilistic predictions are spatially invariant. NM acts as a reference compared to the other models.

2.3.2. *Quantile regression forest*

Quantile regression forest (QRF) (Meinshausen, 2006) is a probabilistic adaption of the random forest (RF) algorithm (Breiman, 2001). QRF and RF make use of an ensemble of decision trees. Each single decision tree of the ensemble is grown with a recursive partitioning of the feature space on an individual bootstrapped training dataset, in which different nodes are created. However, only a random subset of covariates is used in the partitioning of each node. The information given at a prediction site is then run through each decision tree to obtain the corresponding terminal nodes, also known as leaves. The RF point prediction is then the weighted mean of the training observations stored in the corresponding leaves of every tree. QRF makes use of the fact that the RF prediction is a linear combination of the training data. It uses the RF weights and indicator transforms of the training data to estimate the CDF at multiple thresholds, from which the quantile prediction is inferred.

QRF was implemented in the statistical language R (R Core Team, 2023) via the *quantregForest* R-package (Meinshausen, 2017). The maximum node size (*nodesize*) and the number of randomly selected covariates in the partitioning of decision trees (*mtry*) were selected on the basis of a grid-search parameter tuning. For the number of trees fitted in the ensemble (*ntree*), we used the default of 500.

2.3.3. *Quantile regression post-processing with a random forest*

Kasraei et al. (2021) introduced quantile regression post-processing (QRPP) to DSM, which originates from the field of hydrology. It makes use of linear quantile regression (QR) (Koenker and Hallock, 2001), which is comparable to standard linear regression but with the

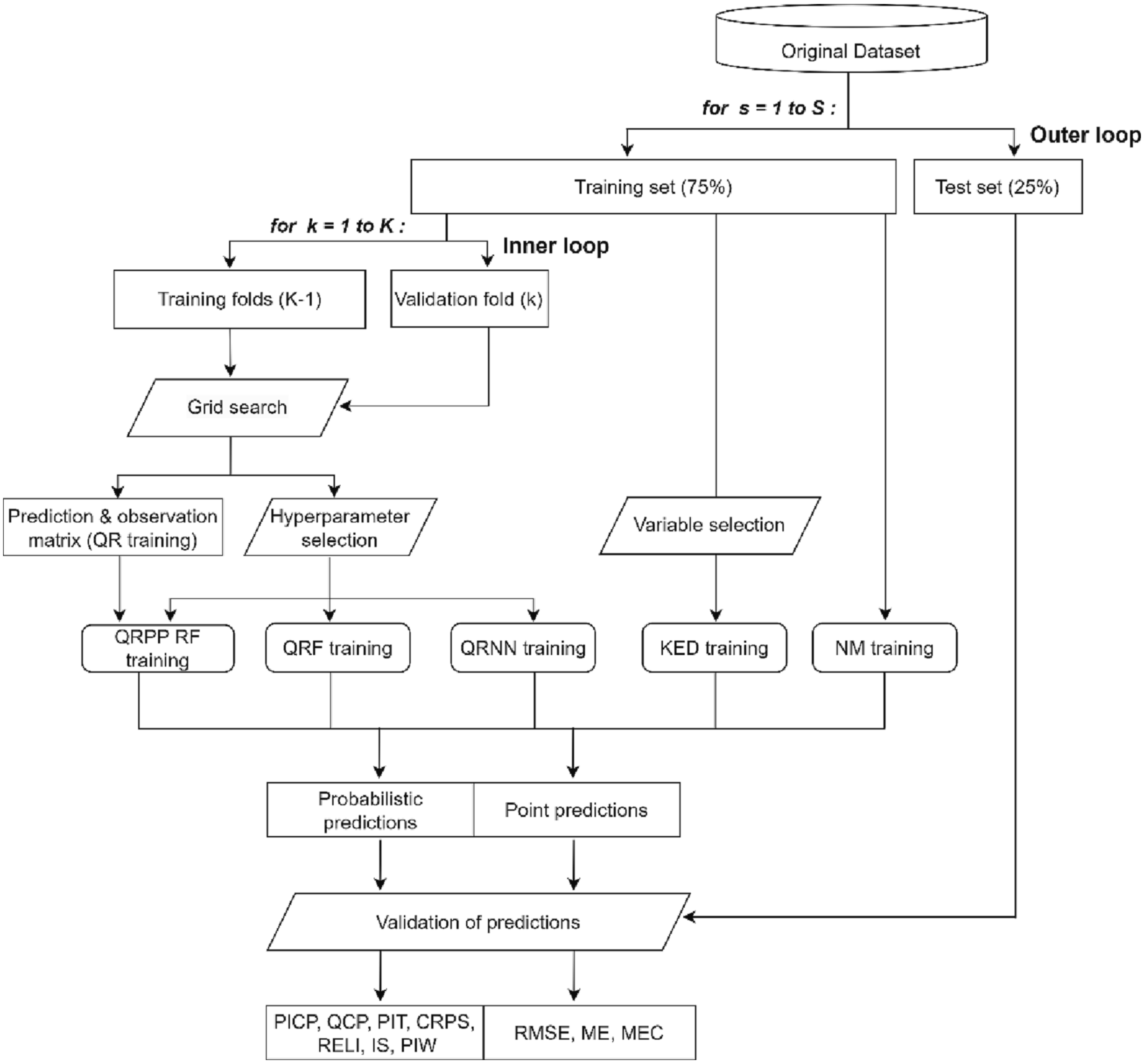

Fig. 2. Conceptualization of the study design.

difference that it predicts conditional quantiles instead of a conditional mean. For that, the quantile loss function, also known as pinball loss function, is minimized in the training process. In QRPP, a QR is fitted on the relationship between point prediction values obtained by a model and observed values. Therefore, the fundamental difference in comparison to other probabilistic prediction models is that the actual probabilistic prediction is not embedded within the model algorithm, making it a two-step procedure. Thus, it is model-agnostic, meaning it can be combined with any point prediction model. In this study, we combined QRPP with an RF model (QRPP RF).

RF was modeled with the *randomForest* R-package (Liaw and Wiener, 2022). The parameters for RF were selected with the same parameter tuning procedure as in QRF. QR was implemented through the *quantreg* R-package (Koenker, 2022).

2.3.4. Kriging with external drift

Kriging with external drift (KED) (Webster and Oliver, 2007) is a hybrid interpolation technique, based on a geostatistical model that represents the dependent variable as the sum of a non-constant trend, i.e. external drift, and a zero-mean stochastic residual. The external drift is usually modeled as a linear function of the covariates, while spatial predictions are achieved with a kriging algorithm. The trend parameters are estimated with generalized least squares, in order to account for autocorrelation of the residuals. KED prediction error variances, which are used to generate the predictive distribution, can then be calculated from the KED prediction error variance, which accounts both for trend estimation errors and residual prediction error (Brus and Heuvelink, 2007). Finally, it is assumed that the predictive distribution follows a normal distribution (Goovaerts, 2001; Heuvelink, 2018).

For each training set, a stepwise-variable selection based on the Akaike Information Criterion was implemented. The variograms needed for kriging were fitted with the *automap* R-package (Hiemstra, 2022), and KED was executed with the *gstat* R-package (Pebesma, 2022).

2.3.5. Quantile regression neural network

Quantile regression neural networks (QRNN) (Cannon, 2011) is a probabilistic adaption of an artificial neural network (ANN). QRNN has the classic multilayer perceptron architecture, which consists of multiple layers, including an input layer, at least one hidden layer and an output layer. Layers are composed of neurons that are connected to the neurons of the previous and following layer. These connections have an associated weight term and each neuron possesses a bias term, apart from neurons in the input layer. The neurons of the input layer supply covariate input data to the first hidden layer, in which in every neuron an output is computed from the weights, bias and a defined hidden layer transfer function which introduces non-linearity. The output then serves as input to the next layer. This continues until the output layer is reached, in which the conditional quantiles are computed with an output layer transfer function. The biases and the weights are determined through backpropagation. The loss function used in QRNN is a differentiable approximation of the quantile loss function originating from QR.

QRNN was modeled with the *qrnn* R-package (Cannon, 2019). By default, we used only one hidden layer and the identity function for the output layer transfer function. The number of neurons in the hidden layer (*n.hidden*), the hidden layer transfer function (*Th* and *Th.prime*) and the weight decay (*penalty*) were determined using a hyperparameter selection.

2.4. Estimation of the predictive distribution

A predictive CDF can be generated with either a parametric or a nonparametric approach (Lauret et al., 2019). In a parametric approach, assumptions about the shape of the distribution are made beforehand. Hence, in order to generate a predictive distribution, one first has to determine the desired distribution from a parametric family (e.g. Gaussian, exponential, Weibull) and next estimate the parameters of the predictive CDF. Kriging and hence also KED uses a parametric approach, in which a normal distribution is typically assumed (Section 2.3.4).

With QRNN, QRPP RF and QRF, a non-parametric approach is used based on predicted quantiles. Nonparametric distributions do not have to adhere to restrictive assumptions imposed by the preselected parametric family (Lauret et al., 2019). However, these methods only predict a finite number of quantiles. Hence, if it is desired to generate a predictive CDF from quantile regression methods, one has to approximate the CDF from these quantiles (Zamo and Naveau, 2018). We did this using a quantile set consisting of 199 quantiles at the 0.5% to 99.5% percentile. Additionally, QRF and QRNN did not provide point prediction values directly, so they were obtained by taking the mean of the quantile set.

2.5. Point prediction validation metrics

Next to probabilistic predictions, we also issued and validated point predictions, to show the performance of the models outside of the probabilistic context. The root mean square error (RMSE) is the most commonly used validation measure in DSM (Piikki et al., 2021) and indicates how much the predictions deviate from the observations:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}, \tag{1}$$

where n is the size of the test set, $\widehat{y}_i$ $(i = 1, \cdots, n)$ are the predicted values and y_i the observed values of the test data.

The mean error (ME) is a bias indicator. Other than RMSE, it can have both positive and negative values (Piikki et al., 2021). A value close to zero indicates that the point predictions are free from bias.

$$ME = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i) \tag{2}$$

The Nash–Sutcliffe model efficiency coefficient (MEC) (Nash and Sutcliffe, 1970) is a relative error measure:

$$MEC = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{3}$$

in which $\overline{y}$ is the arithmetic mean of the test data. In case of perfect agreement between test observations and predictions, the MEC is equal to one. The NM is expected to have a MEC close to zero. Note that MEC will be negative for models that perform worse than the NM.

2.6. Uncertainty validation metrics

2.6.1. Prediction interval width

The prediction interval width (PIW) is a measure for the sharpness of a probabilistic prediction. PIW indicates the width of a certain $\tau \bullet 100$ per cent PI, for any value of τ between 0 and 1:

$$PIW(\tau) = \frac{1}{n}\sum_{i=1}^{n}(u_i - l_i), \tag{4}$$

where l_i is the lower bounding quantile and u_i the upper bounding quantile that together define a $\tau \bullet 100$ per cent PI. Usually, central PIs are of interest, meaning that the probability mass below l_i and above u_i are equal. Therefore, l_i and u_i are determined by the chosen τ value:

$$l_i = q^i_{(1-\tau)/2}, u_i = q^i_{(1+\tau)/2}, \tag{5}$$

where $q^i_{(1-\tau)/2}$ and $q^i_{(1+\tau)/2}$ are the $(1-\tau)/2$ and $(1+\tau)/2$ quantiles of the predictive distribution of y_i.

Lower PIW values imply higher sharpness, i.e. lower uncertainty (Kasraei et al., 2021; Pinson et al., 2007). Therefore, lower PIW values are preferred, given the constraints of reliability. The degree of sharpness is also related to the point prediction performance. For example, when we have a small RMSE, then our probabilistic predictions will have high sharpness, given they are reliable. Although PIW formally is not a validation metric, because its value is independent of the test data, it should be included in the evaluation of probabilistic predictions.

2.6.2. Prediction interval coverage probability

To assess the reliability of PIs, PICP is commonly adopted in DSM (Piikki et al., 2021). Most analyses rely on PICP as a single reliability metric (Kasraei et al., 2021; Lagacherie et al., 2019; Szatmári and Pásztor, 2019; Vaysse and Lagacherie, 2017). The underlying idea is to evaluate what percentage of soil samples from the test set lies in the $\tau \bullet 100$ per cent PIs:

$$PICP(\tau) = \frac{1}{n}\sum_{i}^{n}\delta(l_i \leq y_i \leq u_i) \bullet 100, \tag{6}$$

where δ is an indicator function, with a Boolean argument:

$$\delta(t) = \begin{cases} 1 if t is TRUE \\ 0 else \end{cases}. \tag{7}$$

In an ideal case, PICP (τ) is equal to $\tau \bullet 100$ per cent, e.g., for a 90% PI we desire a PICP of 90%. Multiple PICPs are usually calculated for different PI levels (τ values). This then represents the reliability over the whole predictive distribution. A reliability plot allows a visual evaluation of the reliability by plotting the PICP against the associated PI level. It is then desired that the points are on or close to the 1:1 line. Values below and above the 1:1 line indicate over-pessimistic or over-optimistic PIs, respectively. Note, that in DSM the reliability plot was introduced in Goovaerts (2001) and referred to as 'accuracy plot'. However, 'reliability plot' is a more generally accepted term within other academic fields (Lauret et al., 2019). Schematic examples of PICP reliability plots are shown in Fig. 3.

One clear advantage of PICP is that its value has an intuitive interpretation. Nonetheless, PICP has also a disadvantage, which has not yet been addressed in DSM. As demonstrated in Pinson and Tastu (2014), PICP does not account for a systematic one-sided bias. This occurs when the quantile predictions of the lower and upper boundary of a PI are both either positively or negatively shifted. For example, for a central 90% PI we expect that 5% of the test data are below the lower boundary and 5% above the upper boundary. However, if we have a one-sided bias, in which both boundaries are shifted by +4%, we would observe that 9% of the test data are below the lower boundary and 1% above the upper boundary. In this case, we would still obtain a PICP of 90%, yet it ignores

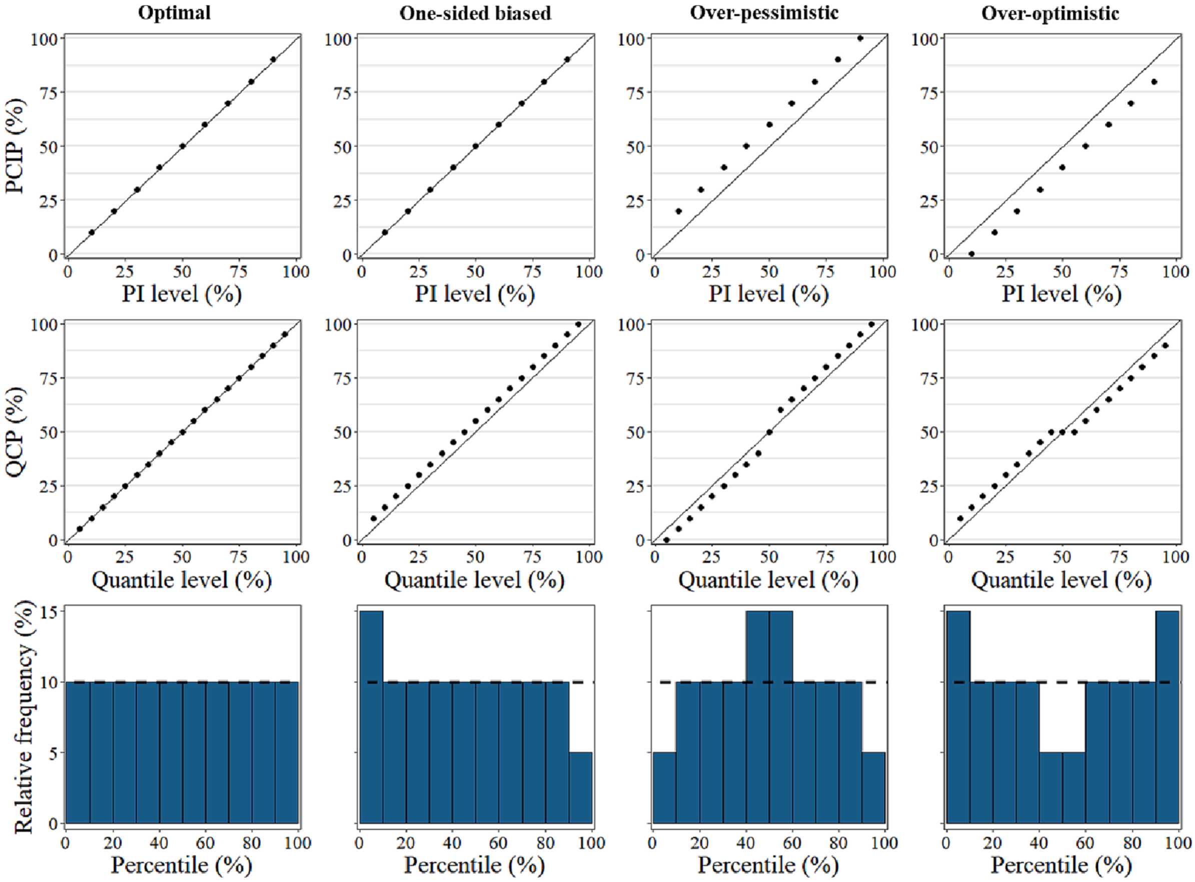

Fig. 3. Reliability plots of PICP and QCP and PIT histograms for four hypothetical scenarios: an optimal, one-sided biased, over-pessimistic and over-optimistic scenario.

the asymmetrical coverage. The effect of one-sided bias on PICP is conceptualized in Fig. 3.

2.6.3. Quantile coverage probability

A simple solution that overcomes the shortcoming of PICP but otherwise has similar properties is the use of the quantile coverage probability (QCP). It has the same underlying logic as PICP but it evaluates single quantile predictions. It computes which fraction of the test set is below a quantile:

$$QCP(\tau) = \frac{1}{n}\sum_{i}^{n}\delta\left(y_i \leq q_\tau^i\right) \bullet 100. \quad (8)$$

This has the advantage that a potential bias will not be hidden. In some studies, only the coverage based on quantiles is computed and the PICP is left out entirely (Lauret et al., 2019; Vasseur and Aznarte, 2021). Examples of QCP reliability plots are also shown in Fig. 3.

2.6.4. Probability integral transform

The probability integral transform (PIT) histogram (Gneiting et al., 2007) is another visual tool for the assessment of reliability. It conveys the same information as QCP but emphasizes different aspects in its visualization. We found one example of PIT usage within DSM in Nussbaum et al. (2014). It evaluates if the test observations y_i cover the whole range of the predictive CDFs. It starts by computing the percentiles P_i associated to the test observations y_i in the predictive distributions F_i:

$$P_i = F_i(y_i). \quad (9)$$

When plotting the obtained P_i as a histogram, this should ideally be a uniform distribution. An uneven distribution indicates that some parts of F_i are disproportionally often or sparsely covered. A sloped, convex or concave shape of the histogram indicates one-sided biased, over-optimistic or over-pessimistic probabilistic predictions, respectively. Over-optimistic refers to a case in which the actual uncertainty is larger than predicted by the probabilistic model, while over-pessimistic refers to the opposite case. All four cases are schematically exemplified in Fig. 3.

2.6.5. Scoring rules

The so far presented validation metrics (PICP, QCP and PIT) indicate the reliability of a single probabilistic prediction model without a required reference or comparison. However, one may also be interested in a relative comparison of competing probabilistic prediction models for model selection or -optimization (Gneiting and Raftery, 2007). For this purpose, scoring rules can be used. Scoring rules are measures that evaluate the quality of a probabilistic prediction model and return a numeric score value. Based on the obtained score values the performance of competing probabilistic prediction models can be compared and ranked. Further, it is desired that scoring rules are proper. The term proper refers to the concept that there is no incentive to report any predictive distribution other than the one of true belief of the model (Gneiting and Raftery, 2007). In the next subsections, we suggest two proper scoring rules. Both are negatively oriented, so that smaller values indicate a better score. Further, they consider both sharpness and reliability.

2.6.6. Interval score

The Interval Score (IS) is a scoring rule that evaluates PIs (Bracher et al., 2021; Gneiting and Raftery, 2007). Therefore, IS depends on the chosen τ value. It is calculated for each test observation y_i and subsequently averaged over the whole test set, to get one final score value:

$$IS(\tau) = \frac{1}{n}\sum_{i=1}^{n}(u_i - l_i) + \frac{2}{1-\tau}\bullet\frac{1}{n}\sum_{i=1}^{n}(l_i - y_i)\bullet\delta(y_i < l_i) + \frac{2}{1-\tau}\bullet\frac{1}{n}\sum_{i=1}^{n}(y_i - u_i)\bullet\delta(y_i > u_i). \quad (10)$$

The first term in Eq. (10) is the average width of the PIs, meaning that sharper PIs receive a lower penalty. The other two terms only consider test observations that are below l_i and above u_i. These observations get a penalty with the distance to the boundaries of the PI. Hence, unlike PICP and QCP, IS also penalizes how far outside a PI the observations are. This may seem contradictory at first because unless we issue a 100% PI, it is desired that a fraction $1-\tau$ of the observations are outside the PI. Yet, as already mentioned, IS additionally considers the width of a PI in its scoring. Therefore, a wider PI may lead to fewer observations outside its boundaries but it simultaneously gets punished for its low sharpness.

To our knowledge, IS was not yet used in DSM. The fact that IS is a scoring rule that evaluates probabilistic predictions in PI format can be an advantage (Bracher et al., 2021) since probabilistic predictions in DSM are usually issued as PIs. Fig. 4 shows a schematic visualization of how single IS(τ) values are calculated.

2.6.7. *Continuous ranked probability score*

The Continuous Ranked Probability Score (CRPS) is a widely used scoring rule for continuous variables (Lauret et al., 2019). We know of two instances where CRPS was used in DSM (Caubet et al., 2019; Nussbaum et al., 2014). Other than IS, CRPS directly evaluates the whole predictive CDF. The calculation of CRPS is comparable to that of point prediction metrics like the mean squared error. It is defined as the integral of the squared difference between the predictive CDFs F_i and empirical CDFs from observed test data. The latter can be also interpreted as a Heaviside step function H, since its CDFs are generated from single test samples y_i:

$$CRPS = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}(F_i(y) - H(y - y_i))^2 dy, \quad (11)$$

where

$$H(t) = \begin{cases} 1 if t \geq 0 \\ 0 else \end{cases}. \quad (12)$$

A schematic visualization of how a single CRPS value is calculated is given in Fig. 4. Additionally, the median CRPS can be evaluated to reduce the influence of outliers in the scoring.

For probabilistic models that return quantiles, there is no continuous predictive CDF as required for Eq. (11). Instead, a step function is generated from a quantile set which approximates the continuous CDF (Section 2.4). In this context, the equations from Hersbach (2000) can be applied, see SI. To calculate CRPS, *crpsDecomposition* from the *verification* R-package (NCAR - Research Applications Laboratory, 2015) was used. This package applies an equation from Hersbach (2000), i.e. Eq. A1 in SI. We also used it in the case of KED, even though KED provides a continuous predictive CDF. However, in case of KED, Eq. A1 functions as numerical integration and the approximation error in comparison to Eq. (11) will be small due to the large number of quantiles used in the approximation.

2.6.8. *Reliability decomposition*

CRPS can be decomposed into different parts (Hersbach, 2000). In this study, we only introduce and use the reliability part (RELI). In RELI,

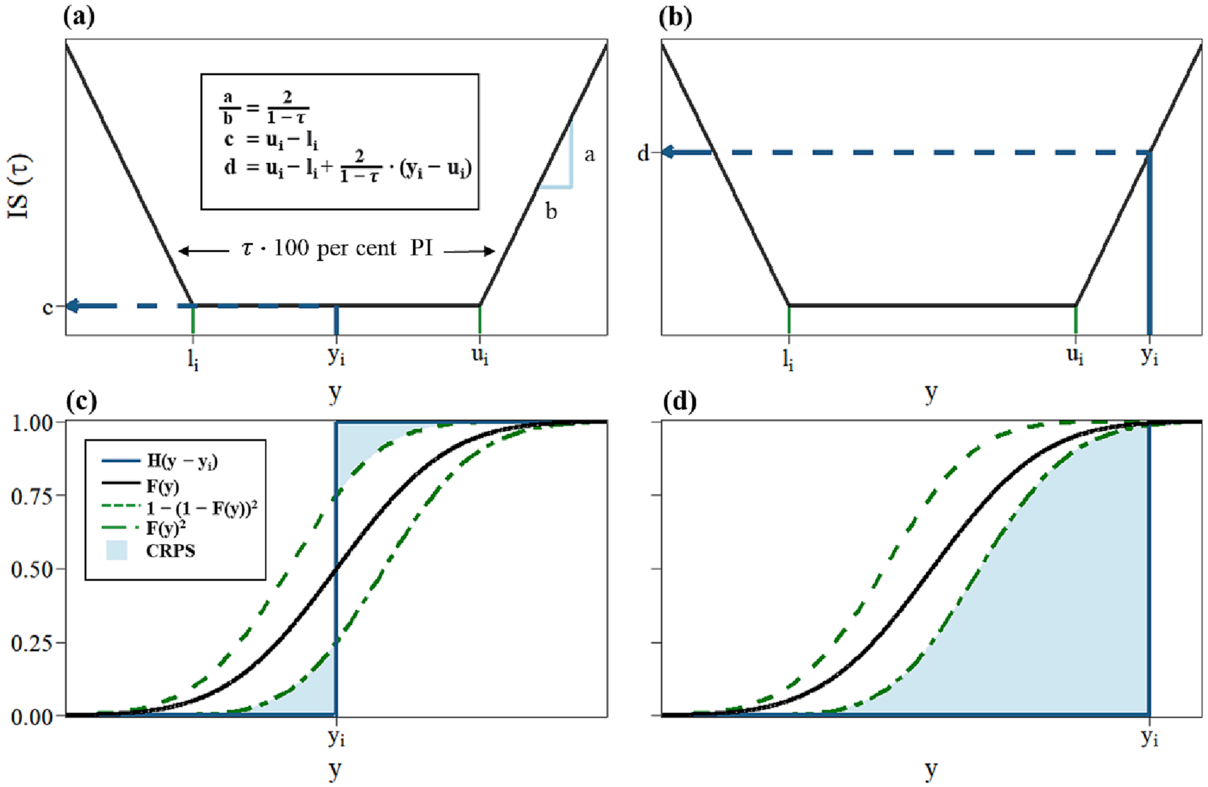

Fig. 4. Schematic representation of two examples of how individual values of IS(τ) (a-b) and CRPS (c-d) are calculated. The examples on the left show a case in which the mean of the predictive distribution is the same as the observed value; the examples on the right show a predictive distribution where the observed value is at the extreme of the distribution. Note, that the area of CRPS is squared, hence the added $1-(1-F(y))^2$ and $F(y)^2$. This figure was inspired by illustrations in Bracher et al. (2021).

the mean coverage of the quantiles used to approximate the predictive CDFs are evaluated. Therefore, it is closely related to QCP and PIT but it returns a single numerical value. It further considers the distance between the quantiles in its weighting. More technical information about the computation of RELI is given in SI. As for CRPS, RELI was computed by the function *crpsDecomposition* from the *verification* R-package.

3. Results

In the following sections, only figures for pH are shown. Figures for log(SOC) can be found in SI. We refer to both soil properties in the text but prioritize pH.

3.1. Point prediction performance

According to MEC and RMSE presented in Table 1, the point prediction performances of pH from QRF, QRPP RF, KED and QRNN were very similar. KED had the single best RMSE of 0.81 but QRPP RF and QRF were close with an RMSE of 0.82 and QRNN with 0.83. As expected, by far the worst point predictions were produced by NM, whose RMSE was with 1.26 around 35% bigger than that of the other models. QRF, QRPP RF and QRNN obtained negative ME values that deviated most from 0. Nevertheless, with a ME of − 0.013, −0.015 and −0.017 respectively, the differences to zero were very small compared to RMSE which indicates that systematic prediction errors were negligible.

For log(SOC), point prediction performances were considerably different (Table A1 in SI). Here, QRF and QRPP RF generated the best results and slightly outperformed QRNN. KED was apart from NM the worst model. Overall, with a maximum MEC of 0.43, log(SOC) point predictions were poorer compared to pH, where the highest MEC was 0.58.

3.2. Probabilistic prediction performance

3.2.1. Prediction interval width

Fig. 5 illustrates PIW for pH at various PI levels to indicate the sharpness of the five prediction models: NM, QRF, QRPP RF, KED and QRNN. Throughout the predictive distribution, NM received the highest PIW values, which were on average 36% larger compared to the other models. KED and QRPP RF shared similar PIW values. For PI between 10% and 90%, PIW of QRF was slightly larger than those of KED and QRPP RF but at the extremes, their values had a similar level. Also, QRNN had similar PIW values as KED and QRPP RF up to the 60% PI. Thereafter, QRNN obtained PIW values that were much smaller compared to the other models. For example, for the 99% PI, the PIW of QRNN was about 46% smaller than that of QRPP RF, which had the second lowest value. Almost the same patterns were found for log(SOC) (Fig. A2), except for KED. Here, KED was less sharp than QRPP RF and had PIW values that were similar to that of QRF.

3.2.2. Prediction interval coverage probability & quantile coverage probability

Fig. 6 and Fig. 7 show PICP and QCP reliability plots of pH, respectively. The evaluated quantiles in Fig. 7 correspond to the PIs in Fig. 6. The PICP and QCP values of NM, QRPP RF and KED were close to the 1:1 line, which is an indicator of good reliability. PICP values of QRF were fairly over-pessimistic, so that in some instances its PICP was around 5% above the 1:1 line. Yet, good concordance was found at the extremes, more specifically above the 90% PI and below the 10% PI. This corresponds to good agreement in terms of QCP at the extremes, i.e. below the 5% quantile and above the 95% quantile, and in the center, i.e., between the 45% and 55% quantile. Between the 5% to 45% quantile, QCP of QRF tended to be below and between the 55 to 95% above the 1:1 line. Since PICP combines the deviation from its lower and upper boundaries the deviation was more visible for QRF in Fig. 6 compared to Fig. 7. QRNN had a very different outcome compared to the other models. In regards to PICP, it performed well until the 60% PI. This corresponds to good agreement in terms of QCP between the 20% and 80% quantile. However, QCP at the edges and PICP for the large PIs showed strong deviations from the 1:1 line. For instance, at the 99% PI, PICP was about 20% below the 1:1 line. There was no meaningful one-sided bias in any model for probabilistic predictions of pH according to the reliability plot of QCP (Fig. 7).

For log(SOC), the trends found for PICP (Fig. A3) and QCP (Fig. A4) of NM, QRPP RF and QRF were similar to those of pH. In contrast, KED obtained over-pessimistic results that were comparable to QRF, judging based on PICP. However, when looking at QCP, additionally one-sided bias was found for KED, so that the deviation from the 1:1 line was more pronounced for KED than for QRF (Fig. A4). In the range between the 30% to 85% quantile, QCP of KED was systematically above the 1:1 line. For example, the PICP corresponding to the 10% PI was around 11%. Yet, for the corresponding 45% and 55% quantile, a QCP of around 49% and 60% was achieved, respectively. QRNN also showed one-sided bias between the 30% and 70% quantile but it was less than for KED.

Table 1
Point prediction performances for pH.

	NM	QRF	QRPP RF	KED	QRNN
MEC	−0.0016	0.57	0.58	0.58	0.56
RMSE	1.26	0.82	0.82	0.81	0.83
ME	~0.000	−0.013	−0.015	−0.007	−0.017

3.2.3. Probability integral transform

PIT histograms for pH are provided in Fig. 8. Since the bin width was 10%, for good reliability it is desirable that the frequency in each bin is close to 10%, as indicated by the horizontal dashed lines in Fig. 8. This was more or less achieved for NM, QRPP RF and KED. In these cases, the relative frequency neither exceeded 11% nor fell below 9%. A concave histogram was obtained for QRF, meaning that lower relative frequency values were obtained at the edges, i.e. the 0% to 10% and 90% to 100% percentile range. Such a shape is characteristic for an over-pessimistic performance. A somewhat convex distribution was achieved for QRNN. Here, the relative frequencies at the edges were around 15%. On the other hand, the bins before and after the edges (10% to 30% and 70% to 90% percentiles), had very diminished relative frequencies.

The PIT histograms obtained for log(SOC) (Fig. A5), were similar to those of pH, except for KED. What stands out is that the PIT histogram of KED was the only histogram without any symmetrical structure. Between the 10% and 90% percentile, frequencies started at a high level above the 10% line but decreased steadily so that after the 60% percentile values were below 10%. Contrary to that trend, the edges did not reflect the same behavior. For the 0% to 10% bin, a decreased frequency was obtained. For the 90% to 100% percentile, the frequency was slightly above 10%.

3.2.4. Interval score

Fig. 9 shows the IS over the whole predictive distribution of pH. Between the 1% and 90% PI, the same ranking order can be found: NM obtained the largest and thus worst scores. The IS of the other models were overall similar but the best scores were obtained for QRF, followed by QRPP RF, KED and lastly QRNN. With increasing PI levels, all acquired IS values of the models rose. Yet, IS of QRNN and KED rose disproportionally more. As a consequence, QRNN and KED surpassed NM at the 99% PI, at which the score of QRNN overshadowed all other models. For log(SOC) (Fig. A6), more or less the same trends were found as for pH, with the only difference that KED was slightly worse than QRNN up to the 80% PI. Just thereafter, QRNN scored worse.

3.2.5. Continuous ranked probability score & reliability decomposition

RELI, CRPS and median CRPS for pH are given in Table 2. All single CRPS values (25 × 2016) obtained in the outer loop are shown as

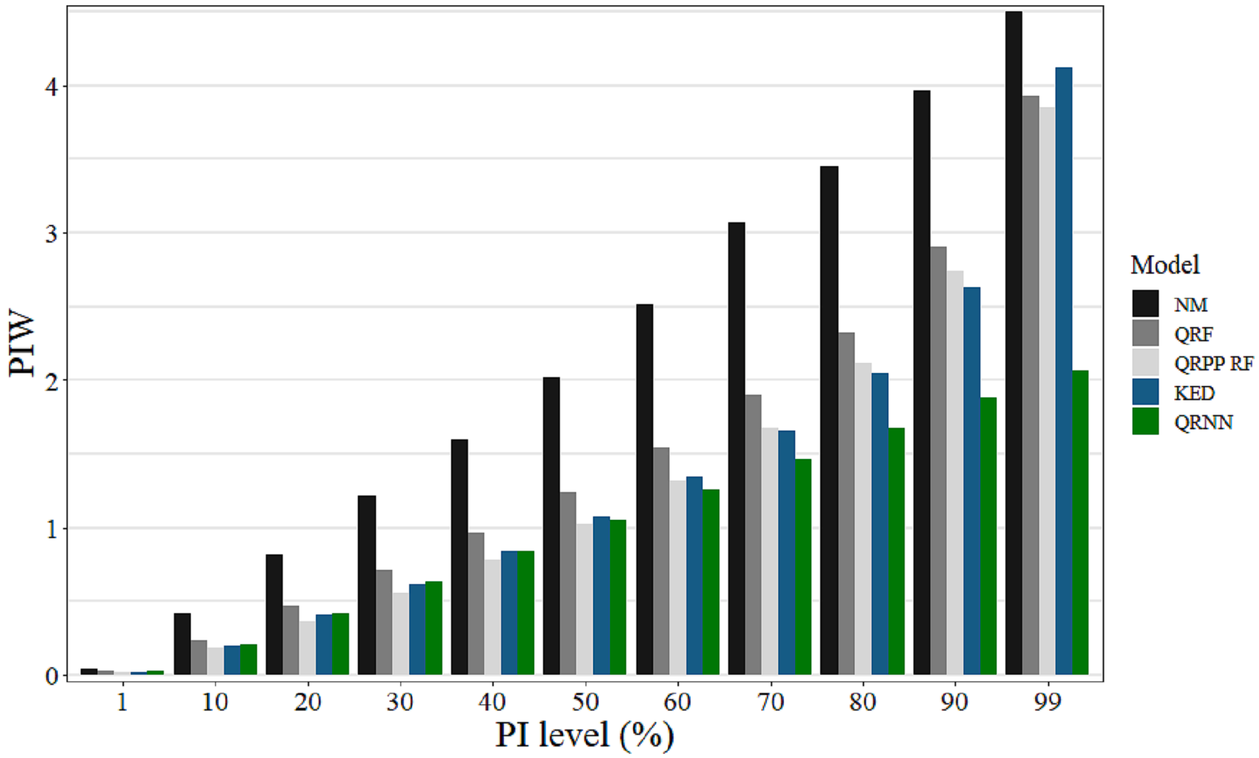

Fig. 5. Sharpness diagram indicating the PIW for multiple PI levels of the predictive distribution for pH.

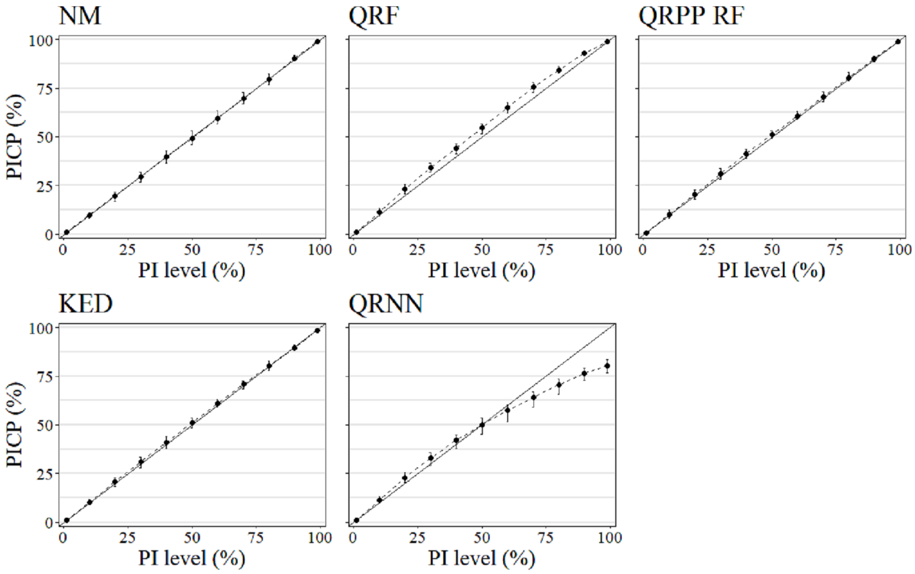

Fig. 6. Mean PICP reliability plots for pH. Error bars are retrieved from the 80% confidence interval of the 25 repetitions in the outer loop. The 1:1 black line indicates the desired outcome.

boxplots in Fig. 10. CRPS was the lowest and thus the best for QRF and QRPP RF. The performance was followed by KED, QRNN and lastly NM. This indicates that based on CRPS, QRF and QRPP RF were able to obtain better uncertainty predictions for pH than KED, QRNN and NM. Yet, the median CRPS of QRF was worse than that of KED and QRNN (Table 2) but the spread of KED and QRNN seemed to be larger (Fig. 10). The ranking order of RELI was very different compared to the ranking of CRPS and median CRPS, as RELI only evaluates reliability. According to RELI, the best reliability was achieved with KED, QRPP RF and NM. RELI of QRF was considerably worse and QRNN scored by far the worst.

For log(SOC), the ranking order with respect to CRPS was slightly different compared to pH (Table A2). Here, QRF and QRPP RF again scored the best but this time, QRNN scored better than KED. KED was also worse in terms of median CRPS. The spread of CRPS was very similar between the models (Fig. A7). Furthermore, KED achieved no longer the best reliability according to RELI. Better results were obtained by NM, QRPP RF and QRF but KED still achieved a lower RELI value than QRNN.

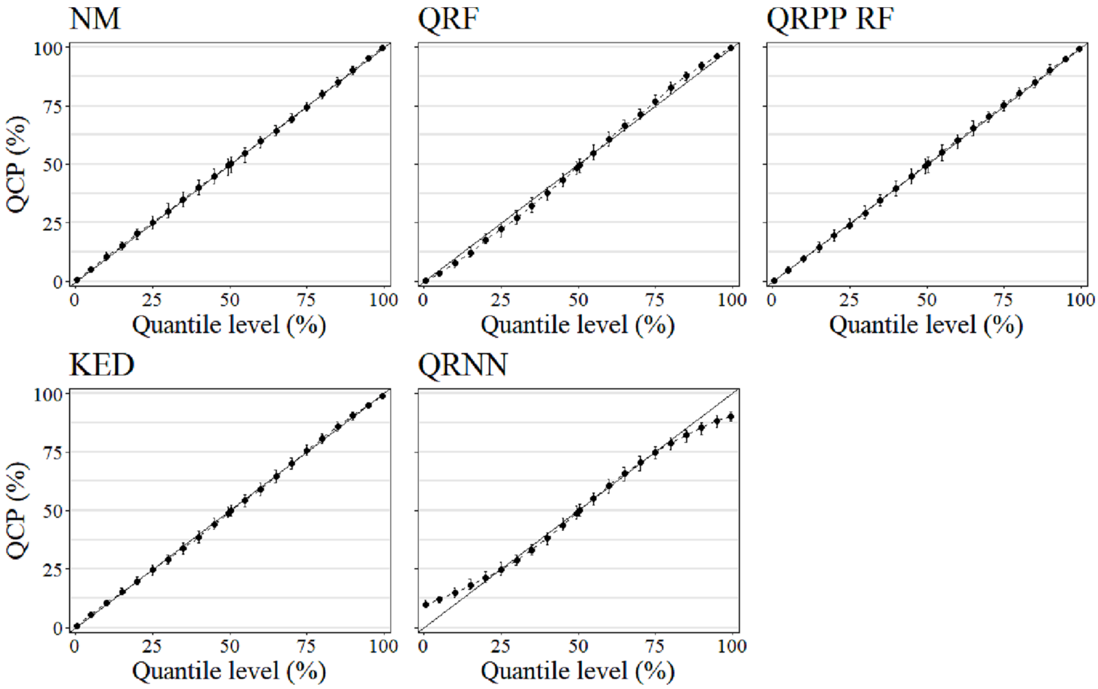

Fig. 7. Mean QCP reliability plots. Error bars are retrieved from the 80% confidence interval of the 25 repetitions in the outer loop. The 1:1 black line indicates the desired outcome.

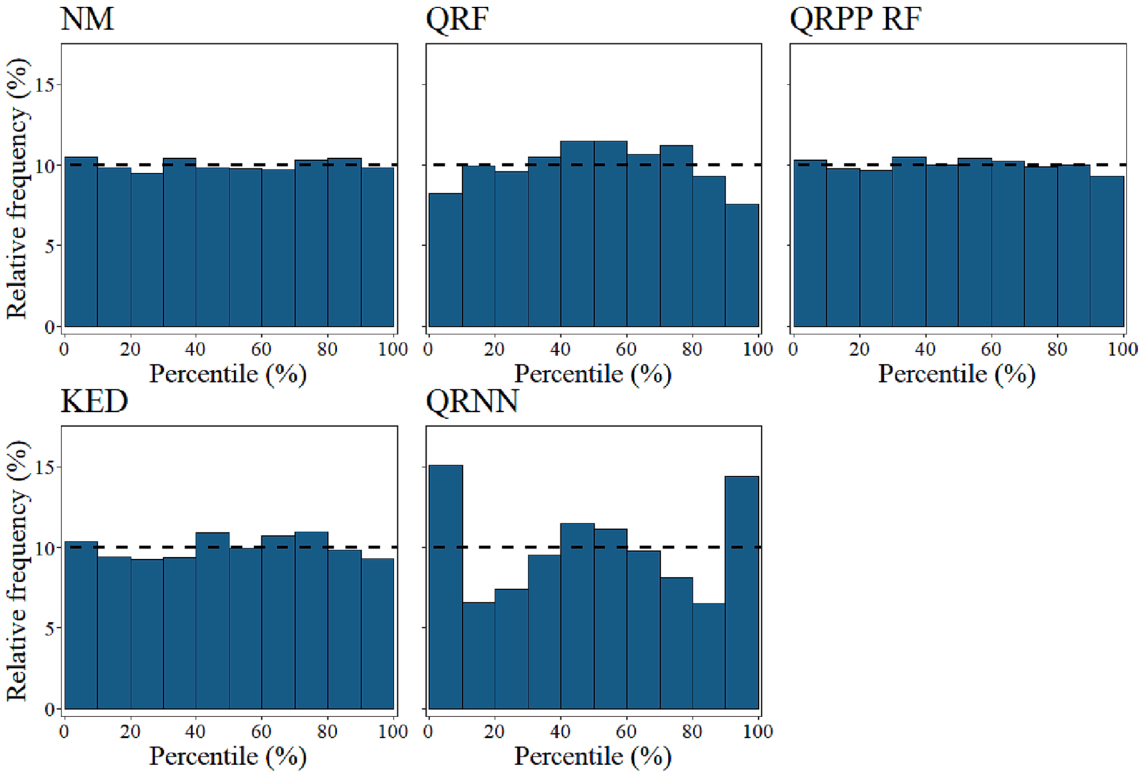

Fig. 8. PIT histograms of pH. The dashed line indicates the desired frequency for a flat and uniform PIT.

4. Discussion

4.1. Comparison of model performance

4.1.1. Point prediction performance

NM used the mean value of the training set as point predictions. Therefore, it is not surprising that it delivered the worst point prediction performance, both for pH and log(SOC). QRF and QRPP RF provided stable point prediction results. This is in line with findings of other DSM studies, where good performances of the associated RF were reported (Khaledian and Miller, 2020). In the case of pH, QRF and QRPP RF were on par with - or slightly worse - than KED. It thus appears that strong

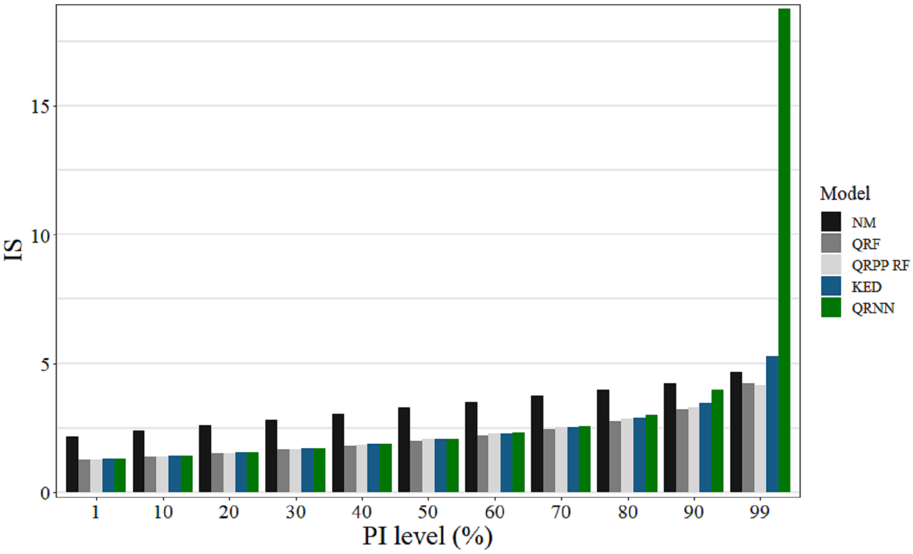

Fig. 9. IS diagram of pH indicating the IS values for multiple PI levels of the predictive distribution.

Table 2
Scoring outcomes of CRPS, median CRPS and RELI for pH.

	NM	QRF	QRPP RF	KED	QRNN
CRPS	0.72	0.44	0.44	0.46	0.47
Median CRPS	0.61	0.33	0.31	0.32	0.32
RELI	0.0016	0.0071	0.0012	0.0010	0.0117

linear relationships between pH and the available covariates were present next to some degree of spatial autocorrelation. Contrarily, nonlinear relationships or interactions seemed to be present between log (SOC) and the covariates because QRPP RF, QRF and also QRNN outperformed KED. There was a small mismatch between QRPP RF and QRF point prediction results, even though both models in theory should achieve a similar outcome. This might be the result of different implementations in the R-packages or small errors with QRF at the conversion of quantile predictions to point predictions. However, the mismatch was very small and insignificant.

4.1.2. Probabilistic prediction performance

It is no surprise that NM provided very reliable results since the uncertainty was modeled by the empirical distribution of the training set. Nonetheless, it suffered from very low sharpness as can be seen from PIW. Due to its inability to issue probabilistic predictions with high sharpness, it was ranked last according to CRPS and IS.

QRF is a very commonly used method to estimate the uncertainty of an RF in DSM. Nonetheless, slightly over-optimistic probabilistic predictions were found in the center of the distribution. This seems to be a common outcome for QRF, as similar problems were reported in Kasraei et al. (2021), Szatmári and Pásztor (2019) and Vaysse and Lagacherie (2017), or outside the DSM literature in for example Vasseur and Aznarte (2021). Yet, according to CRPS and IS, QRF achieved along with

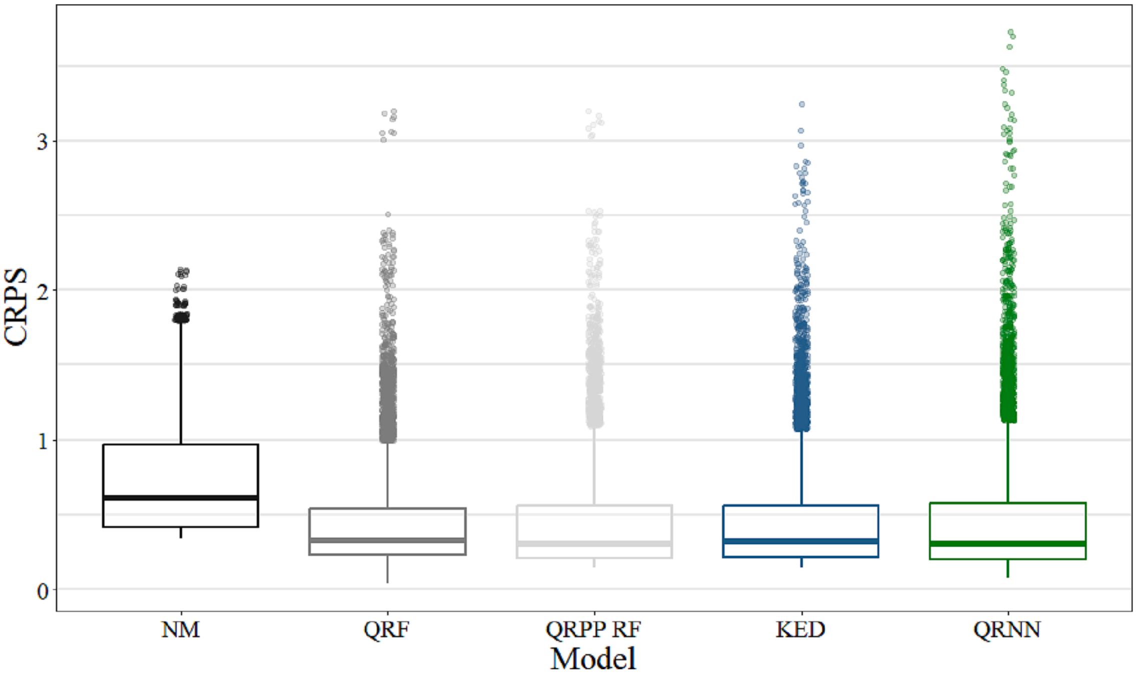

Fig. 10. Single CRPS values portrayed as boxplots for pH. Each boxplot was based on 25 × 2016 values.

QRPP RF the best probabilistic prediction performances. The reasons for QRF performing well compared to other models despite having slightly over-pessimistic probabilistic predictions are discussed later in this section.

QRPP RF performed most consistently on all validation metrics for log(SOC) and pH. The good performance is in agreement with the reported PICP values in Kasraei et al. (2021). This outcome of QRPP RF may be surprising because it uses a simple structure, centered around a linear QR fitted on predicted and observed values. The method of QRPP strongly depends on the residual structure, so in future studies it could be investigated in what way the residual structure influences the probabilistic prediction performance.

Probabilistic predictions of KED were inconsistent in terms of reliability. Good reliability and sharp distributions were found for pH, whereas for log(SOC) considerably worse reliability and unsharp distributions were obtained that additionally had a one-sided bias. Suboptimal probabilistic predictions with different forms of kriging were earlier reported in Vaysse and Lagacherie (2017) and Szatmári and Pásztor (2019). KED had worse CRPS and IS values compared to QRF and QRPP RF, and also QRNN in case of log(SOC). For pH, this may seem counterintuitive at first. Here, KED was sharper and more reliable than QRF. Therefore, one might expect KED to obtain a better score. However, CRPS and IS are sensitive to outliers. This is specifically reflected by IS at the 99% PI, i.e. IS(0.99) (Fig. 9). For IS(0.99), KED scored even worse than NM, despite KED having better sharpness and a PICP close to 99%. When evaluating how IS is calculated (Eq. (10), it is clear that IS imposes a linear penalty depending on how far outside the PI boundaries the test observations fall. Consequently, in the case of KED, the test observations that were not within the 99% PI had a large distance to the boundaries of the PI. It is also noteworthy that KED performed better than QRF in terms of the median CRPS. This is due to the fact that the median CRPS ignores issues with spread and outliers.

QRNN obtained strongly over-optimistic probabilistic predictions. Therefore, QRNN had the worst outcomes among all reliability measures and was also apart from NM the worst with respect to the scoring rules. We do not know what caused these issues. A possible explanation might be that the hyperparameter selection of QRNN was based on its point prediction performance and that the selected hyperparameters influenced the probabilistic prediction performance. The relationship between hyperparameters and probabilistic performance might need more caution considering that potential issues with overfitting of QRNN were discussed in Zhang et al. (2019). Nonetheless, it has to be noted that in one example outside DSM, decent reliability was achieved with QRNN using the same R-package (David et al., 2018). Therefore, we encourage more research with QRNN, in which its application is tested and if needed optimized. Furthermore, there are other noteworthy probabilistic adaptions of neural networks that are popular and eventually useful for DSM. Two examples are the lower upper bound estimation (LUBE) neural network, which predicts PIs (Khosravi et al., 2011), and the ‘improved’ version of QRNN (iQRNN), which aims to prevent overfitting and is reportedly faster (Zhang et al., 2019).

4.2. Value of the proposed uncertainty metrics

Developing tools that help end users to incorporate uncertainty in decision making was identified as one major challenge in pedometrics (Wadoux et al., 2021), which prompted increasing interest in this topic (Breure et al., 2022; Lark et al., 2022). However, the validation of probabilistic predictions has not yet received enough attention, even though a correct validation framework is necessary in order to verify if the uncertainty was quantified reliably and fit for further usage. So far, most analyses relied entirely on PICP when validating the reliability of probabilistic predictions. However, as introduced in the methodological framework but also now shown in a real-world case with KED and QRNN for log(SOC), PICP can hide one-sided bias of the estimated quantiles that set the boundaries of a PI. In these instances, PICP did not capture the actual probabilistic performance well and indicated better results than actually present. One-sided bias may not occur frequently in practice as the other models did not show the same issues. Nonetheless, in order to fully capture the reliability of probabilistic predictions, PICP should preferably be accompanied by a metric that can compensate for that weakness, such as the PIT histogram and QCP. Additionally, the PIT histogram and QCP have an intuitive interpretation similar to that of PICP. Therefore, we strongly encourage adopting these metrics whenever the reliability of an uncertainty map is validated in DSM, so that they either supplement or replace PICP.

Scoring rules allow ranking the probabilistic performance of the models based on a returned numeric score value. RELI solely evaluated the reliability of probabilistic predictions and thus mostly reflected the trends found with PICP, QCP and PIT histograms. Hence, RELI is a useful metric to summarize probabilistic prediction performances with regard to reliability. CRPS and IS were more sensitive to outliers and sharpness. Since sharpness and point prediction performances are related, CRPS and IS reflect trends observed with point prediction validation metrics. This was especially apparent for NM, which scored last due to its low sharpness, explained by the poor point prediction performance. Furthermore, the strong effect of outliers on IS and CRPS may be seen as a nuisance because it means that the outcome can be influenced by a few bad predictions. On the other hand, if test observations are too often found in areas of the predictive CDF that had a low probability density, it can confidently be interpreted as a major flaw of the probabilistic prediction model. However, at least for CRPS, the effect of outliers could be removed by evaluation of the median CRPS. For IS, the returned score additionally depended on the PI level. The penalty for observations outside the PI was more severe with increasing τ, i.e. it was bigger at larger PI levels. Hence, the final judgment should not be grounded on one PI level evaluated by IS.

Usually, when we validate a single PIW-uncertainty map in DSM, we are mainly interested in the reliability of the uncertainty map. In such case, QCP and PIT histograms next to PICP are preferred metrics, as they have an intuitive interpretation. They can inherently show if an uncertainty map is reliable and thus safe to use for an end-user. One can also provide numerical scores from scoring rules but there might not be a great benefit in doing so, because scoring rules are mainly useful for the purpose of ranking model performance. Users are not necessarily interested in comparing performances of models. Researchers on the other hand are more often interested in ranking competing probabilistic prediction models (e.g. Kasraei et al., 2021; Szatmári and Pásztor, 2019; Vaysse and Lagacherie, 2017). For such purposes, scoring rules are a very useful tool by which new information about the weaknesses and strengths of models may become apparent.

4.3. Beyond the proposed metrics

As demonstrated and shown in this study, the validation of probabilistic predictions is more complicated than for point predictions. In this paper, we proposed metrics from the broader probabilistic literature (e.g. Bracher et al., 2021; Brown et al., 2010; Gneiting et al., 2007; Gneiting and Raftery, 2007; Lauret et al., 2019; Pinson et al., 2007) that we deemed useful, intuitive and easy to implement for the context of DSM. Nonetheless, a short outlook is given about concepts and metrics that we did not address but may be worth exploring.

There is a large pool of other scoring rules which can be used to validate probabilistic predictions (Gneiting and Raftery, 2007). For example, the logarithmic score, also known as ignorance score, is another very popular and commonly used scoring rule. It takes the logarithm of the predictive probability density. Therefore, it is even more sensitive to outliers than CRPS because if the predictive probability density is close to zero, it converges to infinity or minus infinity, depending on whether a negative or positive orientation is chosen. However, Bracher et al. (2021) argued that the edges of a probability density function may not be reliably approximated from a set of predicted quantiles.

In Section 4.2 we stated that it may not be necessary to include scoring rules when validating an uncertainty map. However, in other academic fields, scoring rules are sometimes expressed in form of skill scores (Gneiting and Raftery, 2007; Lauret et al., 2019). A skill score measures the relative performance, where an obtained score is compared to a reference, for which NM is a natural choice. As such it has a similar logic as the MEC for point predictions, for which the NM produces a value close to zero. Therefore, if it is desired to further deliver the probabilistic performance in terms of a scoring rule, a skill score allows for a more general interpretability than the pure absolute numeric scores presented in this study.

The PIT-histograms presented in this study were interpreted upon subjective visual judgment. Goodness-of-fit test could be used to test the 'flatness' of a histogram (Elmore, 2005). It tests the null hypothesis of whether the obtained percentiles used for the PIT histogram follows a uniform distribution. It can be used to reduce the subjectivity of a purely visual assessment.

We restricted our analysis to sharpness and reliability. Yet, there are also other attributes that can be evaluated. For example, in the context of ensemble forecasts, multiple studies advocated to also evaluate probabilistic predictions with the concept of resolution (Brown et al., 2010; Lauret et al., 2019; Pinson et al., 2007). Resolution measures how case-dependent the resulting probabilistic predictions are, meaning that different predictive distributions are generated depending on the covariate condition.

In this study, we only focused on how probabilistic predictions and thus uncertainty maps could be evaluated within DSM with respect to validation metrics. We did not address the importance of a validation strategy (e.g. independent sampling, cross-validation or data-splitting) and sampling design (e.g. probability or non-probability sampling). These are important aspects to obtain unbiased validation results. Yet, they were already studied and discussed in great depth for point predictions (Brus et al., 2011; Piikki et al., 2021) and the same rules apply to the validation of probabilistic predictions. Note that the LUCAS-soil dataset used in this study is based on a multistage stratified random sampling design (Orgiazzi et al., 2018).

5. Conclusion

New metrics and concepts for the validation of probabilistic predictions were introduced and their relevance for DSM studies was illustrated in a case study with five different prediction models for pH and log(SOC). The methodical framework can be used to improve currently used validation procedures in DSM. Our conclusions are:

- PICP lacks the capability to detect one-sided bias in quantile predictions that define the boundaries of PIs. Under such circumstances, PICP fails to reveal issues with reliability. Considerable one-sided bias may occur in practice as shown for KED and QRNN in the case of log(SOC). Therefore, other validation tools like QCP or the PIT histogram should complement or replace PICP.
- Scoring rules such as CRPS, IS and RELI allow for a ranking of probabilistic prediction model performances based on a numeric value. CRPS and IS were sensitive to outliers and sharpness. RELI summarized the trend found with reliability metrics such as PICP, QCP and PIT histograms and was less sensitive to outliers. Yet, it has to be acknowledged that scoring rules are mostly useful for comparing probabilistic performances of competing models.
- Depending on the metrics evaluated, different outcomes can be perceived in terms of probabilistic prediction performance. Therefore, including a set of different validation metrics allows for a more critical evaluation.
- Considering all metrics for evaluating the probabilistic prediction performance of the five prediction models: NM showed high reliability but suffered from low sharpness; QRF had over-pessimistic uncertainty in the center of the distribution but performed well on the edges, QRPP RF was the most consistent; KED obtained inconsistent results and QRNN suffered from low reliability due to over-optimistic probabilistic predictions.

We strongly encourage to use the recommended tools in future studies for a more comprehensive validation of probabilistic predictions in the field of DSM.

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.

Data availability

The authors do not have permission to share the data but the data can be requested by the given data-providers.

Acknowledgements

We thank Dr. James Brown for useful comments and ISRIC - World Soil Information for providing the covariate data.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Supplementary Information

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoderma.2023.116585.

References

Bracher, J., Ray, E.L., Gneiting, T., Reich, N.G., Pitzer, V.E., 2021. Evaluating epidemic forecasts in an interval format. PLoS Comput. Biol. 17 (2), e1008618.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Breure, T.S., Haefele, S.M., Hannam, J.A., Corstanje, R., Webster, R., Moreno-Rojas, S., Milne, A.E., 2022. A loss function to evaluate agricultural decision-making under uncertainty: a case study of soil spectroscopy. Precis. Agric. 23 (4), 1333–1353.

Brown, J.D., Demargne, J., Seo, D.-J., Liu, Y., 2010. The Ensemble Verification System (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. Environ. Model. Softw. 25 (7), 854–872.

Brus, D.J., Heuvelink, G.B.M., 2007. Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138 (1–2), 86–95.

Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. Eur. J. Soil Sci. 62 (3), 394–407.

Cannon, A.J., 2011. Quantile regression neural networks: Implementation in R and application to precipitation downscaling. Comput. Geosci. 37 (9), 1277–1284.

Cannon, A.J., 2019. qrnn: Quantile Regression Neural Network. R-package Version 2, 5.

Caubet, M., Dobarco, M.R., Arrouays, D., Minasny, B., Saby, N.P., 2019. Merging country, continental and global predictions of soil texture: Lessons from ensemble modelling in France. Geoderma 337, 99–110.

Chen, S., Arrouays, D., Leatitia Mulder, V., Poggio, L., Minasny, B., Roudier, P., Libohova, Z., Lagacherie, P., Shi, Z., Hannam, J., Meersmans, J., Richer-de-Forges, A. C., Walter, C., 2022. Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. Geoderma 409, 115567.

David, M., Luis, M.A., Lauret, P., 2018. Comparison of intraday probabilistic forecasting of solar irradiance using only endogenous data. Int. J. Forecast. 34 (3), 529–547.

Elmore, K.L., 2005. Alternatives to the chi-square test for evaluating rank histograms from ensemble forecasts. Wea. Forecasting 20 (5), 789–795.

Gneiting, T., Balabdaoui, F., Raftery, A.E., 2007. Probabilistic forecasts, calibration and sharpness. J Royal Statistical Soc B 69 (2), 243–268.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 102 (477), 359–378.

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

Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Wea. Forecasting 15 (5), 559–570.

Heuvelink, G.B.M., 1998. Error Propagation in Environmental Modelling with GIS. CRC Press.

Heuvelink, G.B.M., 2018. Uncertainty and Uncertainty Propagation in Soil Mapping and Modelling. In: Pedometrics. Springer, Cham, pp. 439–461.

Hiemstra, P.H., 2022. automap: Automatic Interpolation Package. R-package version 1.0-16.

ISO, 1994. Soil quality — Determination of pH. International Organization for Standardization, Geneva.

ISO, 1995. Soil quality — Determination of organic and total carbon after dry combustion (elementary analysis). International Organization for Standardization, Geneva.

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. Environ. Model. Softw. 144, 105139.

Keesstra, S.D., Bouma, J., Wallinga, J., Tittonell, P., Smith, P., Cerdà, A., Montanarella, L., Quinton, J.N., Pachepsky, Y., van der Putten, W.H., Bardgett, R.D., Moolenaar, S., Mol, G., Jansen, B., Fresco, L.O., 2016. The significance of soils and soil science towards realization of the United Nations Sustainable Development Goals. SOIL 2 (2), 111–128.

Khaledian, Y., Miller, B.A., 2020. Selecting appropriate machine learning methods for digital soil mapping. App. Math. Model. 81, 401–418.

Khosravi, A., Nahavandi, S., Creighton, D., Atiya, A.F., 2011. Lower upper bound estimation method for construction of neural network-based prediction intervals. IEEE Trans. Neural Netw. 22 (3), 337–346.

Koenker, R., Hallock, K.F., 2001. Quantile Regression. J. Econ. Perspect. 15 (4), 143–156.

Koenker, R., 2022. quantreg: Quantile Regression. R-package version 5.94.

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., Martin, M., 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.

Lark, R.M., Chagumaira, C., Milne, A.E., 2022. Decisions, uncertainty and spatial information. Spatial Statistics 50, 100619.

Lauret, P., David, M., Pinson, P., 2019. Verification of solar irradiance probabilistic forecasts. Sol. Energy 194, 254–271.

Liaw, A., Wiener, M., 2022. randomForest: Classification and Regression by randomForest. R-package version 4 (7-1), 1.

Malone, B.P., McBratney, A.B., Minasny, B., 2011. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. Geoderma 160 (3–4), 614–626.

McBratney, A., Mendonça Santos, M., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1–2), 3–52.

Meinshausen, N., 2006. Quantile regression forests. J. Mach. Learn. Res. 7 (6).

Meinshausen, N., 2017. quantregForest: Quantile Regression Forests. R-package version 1.3-7.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. J. Hydrol. 10 (3), 282–290.

NCAR - Research Applications Laboratory, 2015. verification: Weather Forecast Verification Utilities. R-package version 1, 42.

Nelson, M.A., Bishop, T.F.A., Triantafilis, J., Odeh, I.O.A., 2011. An error budget for different sources of error in digital soil mapping. Eur. J. Soil Sci. 62 (3), 417–430.

Nussbaum, M., Papritz, A., Baltensweiler, A., Walthert, L., 2014. Estimating soil organic carbon stocks of Swiss forest soils by robust external-drift kriging. Geosci. Model Dev. 7 (3), 1197–1210.

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., Fernández-Ugalde, O., 2018. LUCAS Soil, the largest expandable soil dataset for Europe: a review. Eur. J. Soil Sci. 69 (1), 140–153.

Pebesma, E., 2022. gstat: Spatial and Spatio-Temporal Geostatistical Modelling, Prediction and Simulation. R-package version 2.1-0.

Piikki, K., Wetterlind, J., Söderström, M., Stenberg, B., 2021. Perspectives on validation in digital soil mapping of continuous attributes—A review. Soil Use Manage 37 (1), 7–21.

Pinson, P., Nielsen, H.A., Møller, J.K., Madsen, H., Kariniotakis, G.N., 2007. Non-parametric probabilistic forecasts of wind power: required properties and evaluation. Wind Energ. 10 (6), 497–516.

Pinson, P., Tastu, J., 2014. Discussion of "prediction intervals for short-term wind farm generation forecasts" and "combined nonparametric prediction intervals for wind power generation". IEEE Trans. Sustain. Energy 5 (3), 1019–1020.

Poggio, L., de Sousa, L.M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. SOIL 7 (1), 217–240.

R Core Team, 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

Szatmári, G., Pásztor, L., 2019. Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. Geoderma 337, 1329–1340.

Vasseur, S.P., Aznarte, J.L., 2021. Comparing quantile regression methods for probabilistic forecasting of NO2 pollution levels. Sci. Rep. 11 (1), 11592.

Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma 291, 55–64.

Wadoux, A.M.C., Heuvelink, G.B., Lark, R.M., Lagacherie, P., Bouma, J., Mulder, V.L., Libohova, Z., Yang, L., McBratney, A.B., 2021. Ten challenges for the future of pedometrics. Geoderma 401, 115155.

Webster, R., Oliver, M.A., 2007. Kriging in the Presence of Trend and Factorial Kriging, in: Webster, R., Oliver, M.A. (Eds.), Geostatistics for environmental scientists, Second Edition ed. Statistics in practice. Wiley, Chichester, pp. 195–218.

Zamo, M., Naveau, P., 2018. Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts. Math. Geosci. 50 (2), 209–234.

Zhang, W., Quan, H., Srinivasan, D., 2019. An improved quantile regression neural network for probabilistic load forecasting. IEEE Trans. Smart Grid 10 (4), 4425–4434.

Zhang, Y., Wang, J., Wang, X., 2014. Review on probabilistic forecasting of wind power generation. Renew. Sustain. Energy Rev. 32, 255–270.