The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume L-4/W1-2026
Free and Open Source Software for Geospatial (FOSS4G) 2026 – Academic Track "Bridging Geospatial Technology and Humanity",
30 August–5 September 2026, Hiroshima, Japan

Comparing uncertainty quantification methods for Random Forest-based digital soil mapping

Liina Hints[1], Raivo Aunap[1], Meelis Kull[2], Jeonghwan Choi[1], Alexander Kmoch[1]

[1] Department of Geography, University of Tartu, Tartu, Estonia – (liina.hints, raivo.aunap, jeonghwan.choi, alexander.kmoch)@ut.ee
[2] Institute of Computer Science, University of Tartu, Tartu, Estonia – meelis.kull@ut.ee

Keywords: Soil organic carbon mapping, Uncertainty quantification, Random Forest, Conformal prediction, Area of applicability

Abstract

Uncertainty quantification is essential for using machine learning-based soil maps in decision-making, but the uncertainty communicated to users can depend strongly on the method used. This is especially relevant for soil organic carbon (SOC), where spatial predictions are used in carbon accounting, land management, and climate-related reporting. This study compares five uncertainty quantification methods within a shared Random Forest-based SOC modelling workflow for Estonia: 1) Quantile Regression Forest, 2) Conformalized Quantile Regression, 3) land use-conditional Conformalized Quantile Regression, 4) Random Forest-based calibration of conformity scores, and 5) land use-conditional split conformal prediction. Using 1004 SOC observations and 16 environmental covariates, the study evaluates how these methods differ in coverage, sharpness, land use- and SOC-specific calibration, environmental patterns of high uncertainty, and relationship with Area of Applicability.
All tested methods produced reasonably good marginal coverage for 90% prediction intervals, but differed in their local behaviour and sharpness. The results show that decomposing prediction interval diagnostics by land use and predicted SOC range provides a more informative view of uncertainty quality. High uncertainty was linked to specific environmental covariate combinations, suggesting that interval-based uncertainty patterns could help identify conditions where additional sampling may be useful. Prediction interval width and Area of Applicability were only weakly related. The main contribution of this study is a structured diagnostic framework for comparing prediction interval-based uncertainty methods at both marginal and more local levels, while also improving our understanding of how interval-based uncertainty relates to environmental conditions and model applicability in Estonia's SOC model.

1. Introduction

Machine learning is widely used to derive spatial data for environmental decision-making, including in digital soil mapping (DSM), where mapped products support carbon accounting, land management, crop modelling, food security, and policy reporting. In DSM, soil organic carbon (SOC) has become a key target variable because many climate- and land-related decisions now depend on spatially explicit estimates of SOC stocks and change. Recent advances in machine learning-based spatial modelling have accelerated SOC mapping, while growing standardisation efforts have improved the consistency of data collection and analysis workflows.

Predictions from machine learning-based environmental models are inherently uncertain, both due to incomplete knowledge of the processes controlling the target variable (epistemic uncertainty) and irreducible variability in the system itself (aleatoric uncertainty). Predictions should therefore always be accompanied by their uncertainty estimates to support decision-making and avoid overconfidence in model outputs. Previous studies show that interpretation and decision-making depend not only on the reported uncertainty values, but also on the type of uncertainty reported and how it is communicated. This makes uncertainty quantification especially valuable in open-source applications, as it improves the reliability and explainability of shared models and datasets.

Uncertainty quantification in DSM has received increasing attention in recent years, although most studies either report a single uncertainty measure or compare uncertainty performance across different modelling algorithms. In SOC modelling, uncertainty is most commonly expressed through prediction intervals, often derived from Quantile Regression Forest (QRF) or related ensemble-based approaches, with 90% prediction intervals widely adopted following the GlobalSoilMap framework (Arrouays et al., 2014).

However, even for prediction interval-based methods, the choice of the specific uncertainty quantification method can affect what kind of uncertainty is communicated. Even when the underlying prediction model is the same, different methods may produce intervals that differ in width, calibration, spatial pattern, and local behaviour across land use or SOC ranges. This is important in applied DSM workflows, where the modelling algorithm is often selected first based on predictive performance or practical implementation, and uncertainty quantification is added afterwards.

This study compares five uncertainty quantification methods applicable to Random Forest-based SOC modelling: Quantile Regression Forest, Conformalized Quantile Regression, land use-conditional Conformalized Quantile Regression, random forest-based local conformity-score calibration, and land use-conditional split conformal prediction. The comparison is applied to SOC prediction across Estonia and focuses on how method choice affects the calibration, spatial pattern, and interpretation of the resulting uncertainty estimates. In addition, interval-based uncertainty is compared with the Area of Applicability framework to assess whether wider prediction intervals coincide with areas where the model is applied to less familiar covariate combinations. The aim is to assess how method choice affects the uncertainty that is communicated and whether a combination of multiple methods and diagnostics provides a more complete interpretation of SOC prediction uncertainty.

2. Data and methodology

2.1 Data

This study used 1004 soil organic carbon (SOC) observations from across Estonia (Figure 1). The SOC observations were compiled from existing datasets, including LUCAS Soil (Orgiazzi et al., 2018), EstSoil-EH (Kmoch et al., 2021), and grassland-related SOC data from Helm (2023), together with additional sampling following Choi et al. (2025).

SOC was modelled using 16 environmental covariate rasters representing soil properties, vegetation indices, topographic indices, drainage conditions, and land use. The full list of covariates is shown in Figure 6.

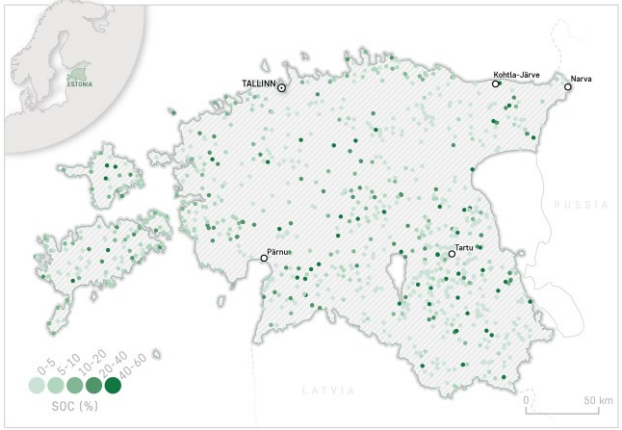

Figure 1. Spatial distribution of SOC samples (n=1004) across Estonia

2.2 Random forest-based uncertainty quantification methods

Five uncertainty quantification methods were compared in this study. The methods were selected because they can be used together with random forest-based modelling and produce a comparable spatial uncertainty expression, either as prediction intervals or as interval-like lower and upper prediction bounds. Quantile Regression Forest (QRF) was used as the baseline probabilistic model, and the remaining methods were applied as calibration or post-processing approaches built around the same underlying prediction problem.

Model fitting and evaluation were implemented in Python using scikit-learn and scikit-learn-compatible estimators (Pedregosa et al., 2011). Model performance and uncertainty diagnostics were evaluated using five-fold cross-validation, so that each observation was predicted out-of-fold. For QRF-based methods, the predictive distribution was queried at quantiles from 0.01 to 0.99, with denser coverage in the tails. These quantiles were used to construct central prediction intervals at several nominal levels and to create reliability plots across the predictive distribution.

2.2.1 Quantile Regression Forest (QRF) (Meinshausen, 2006) was used as the baseline model. QRF extends the random forest framework by estimating conditional quantiles of the response variable, instead of only the conditional mean. This allows the model to return a predictive distribution for each prediction location.

QRF was chosen as the baseline because it is one of the most commonly used machine-learning approaches for uncertainty quantification in digital soil mapping (DSM). Prediction intervals, especially 90% prediction intervals, are also a widely used way of communicating uncertainty in DSM and have been widely adopted following the GlobalSoilMap framework (Arrouays et al., 2014). In this study, QRF provided the reference prediction intervals against which the calibrated and locally adaptive methods were compared.

2.2.2 Conformalized Quantile Regression (CQR) (Romano et al., 2019) was used to recalibrate the QRF prediction intervals. The motivation for using CQR was that QRF can estimate conditional quantiles and produce spatially varying prediction intervals, but the empirical coverage of these intervals may deviate from the nominal level. In other words, a predicted 90% interval does not necessarily contain approximately 90% of independent observations. Conformal prediction addresses this by using calibration residuals to correct the interval bounds after model fitting.

CQR combines these two ideas. The original QRF quantiles provide the shape and local width of the interval, while conformal calibration adjusts the interval so that the desired marginal coverage is reached under exchangeability. Compared with applying conformal calibration directly to point prediction residuals, CQR can thus preserve the locally varying interval structure of the underlying quantile model and has been shown to achieve similar coverage with sharper intervals in many settings (Romano et al., 2019; Zhang et al., 2020).

In this study, the lower and upper QRF quantiles were first predicted for each validation observation. A conformity score was then calculated to measure how far the observed SOC value fell outside the predicted interval:

$$s_i = \max\left(\hat{q}_{\alpha/2}(x_i) - y_i,\ y_i - \hat{q}_{1-\alpha/2}(x_i)\right) \qquad (1)$$

where y_i is the observed SOC value, and $\hat{q}_{\alpha/2}(x_i)$ and $\hat{q}_{1-\alpha/2}(x_i)$ are the lower and upper predicted quantiles. The conformal correction $\hat{q}_{\text{conf}}$ was then obtained from the empirical quantile of the conformity scores and applied to the original QRF interval:

$$C(x) = \left[\hat{q}_{\alpha/2}(x) - \hat{q}_{\text{conf}},\ \hat{q}_{1-\alpha/2}(x) + \hat{q}_{\text{conf}}\right] \qquad (2)$$

This correction is applied globally, meaning that the same adjustment is used for all prediction locations. As a result, CQR targets marginal rather than local coverage. Depending on the calibration errors of the baseline QRF model, the correction can widen or narrow the original intervals. The resulting intervals remain asymmetric because their lower and upper bounds are still based on the original QRF quantile estimates.

2.2.3 Land use-conditional Conformalized Quantile Regression (LU-CQR) was tested as a land use-conditional version of CQR to assess whether class-specific conformal calibration improves coverage within land use classes while preserving the asymmetric prediction intervals inherited from the underlying QRF quantiles. The method follows the same logic as CQR, but the conformal correction was calculated separately for each land use class: arable land, forest, grassland, and wetland. This means that each land use class received its own correction term, instead of being estimated only once for the full calibration set.

While standard CQR targets marginal coverage over the full dataset, this does not necessarily mean that coverage is equally reliable within more specific subsets of the data. Land use was used here as a simple approximation of local variation, because it captures broad differences in SOC-forming conditions and can be mapped spatially.

2.2.4 Random Forest calibration of conformity scores (RFoCS) was implemented using the Loforest method from the clover package, following Cabezas et al. (2025). This method also uses conformity scores, but instead of applying a single global conformal correction, it models the spatial variation of conformity scores as a function of the predictor variables.

First, conformity scores were calculated as absolute residuals around the baseline point prediction:

$$s_i = |y_i - \hat{y}_i| \quad (3)$$

where $\hat{y}_i$ is the baseline prediction for observation i. A random forest was then fitted using the environmental covariates as predictors and the conformity scores as the response. The forest structure was used to estimate local cut-off values for new prediction locations. These cut-offs define how wide the prediction interval should be around the point prediction at each location.

Unlike CQR, this approach produces symmetric intervals around the point prediction:

$$C(x) = [\hat{y}(x) - \hat{t}_{1-\alpha}(x),\ \hat{y}(x) + \hat{t}_{1-\alpha}(x)] \quad (4)$$

where $\hat{t}_{1-\alpha}(x)$ is the locally estimated conformity-score cut-off. The method was included because it is designed to produce locally adaptive interval widths and is therefore relevant when the aim is not only marginal coverage over the full dataset, but also better local representation of uncertainty.

2.2.5 Land use-conditional split conformal prediction (LU-CP) was tested as a simpler group-conditional conformal benchmark. In standard split conformal prediction for regression, one conformity-score cut-off is calculated from calibration residuals and applied to all predictions. This produces one global interval width around the point prediction, which guarantees marginal coverage but does not adapt to local differences in uncertainty.

In LU-CP, the conformity-score cut-off was instead calculated separately for each land use class. Each prediction location then received the interval width corresponding to its land use class. Like RFoCS, this approach produces symmetric intervals around the point prediction. The method was included mainly as an interpretable comparison case, showing how much interval width differs between land use classes when calibration is performed at the class level.

2.3 Prediction interval diagnostics

The methods were compared using several diagnostics that describe different aspects of prediction interval performance. The main focus was on reliability, sharpness, and the balance between these two properties.

Most interval diagnostics describe marginal performance across the full validation dataset. However, good overall coverage does not necessarily mean that uncertainty is equally well represented across different environmental conditions. To gain insight into local reliability and sharpness, the diagnostics were also calculated separately for land use classes (arable land, forest, grassland, and wetland) and predicted SOC concentration classes (0-2.5%, 2.5-5%, 5-10%, 10-40%, and >40%).

2.3.1 Prediction Interval Coverage Probability (PICP) was used to evaluate the reliability of prediction intervals (Shrestha & Solomatine, 2006). PICP is one of the most widely used metrics for assessing the quality of predicted uncertainty and has been used extensively in DSM, especially for 90% prediction intervals, for example in SoilGrids 2.0 (Poggio et al., 2021) and by Kasraei et al. (2021). PICP measures the proportion of observations that fall inside the predicted interval. For a 90% prediction interval, the expected PICP is 0.90, meaning that 90 out of 100 observations should fall inside the predicted uncertainty range. Values below the nominal level indicate that uncertainty is underestimated, while values above the nominal level indicate that uncertainty is overestimated.

PICP was calculated for the 90% prediction intervals of all methods. In addition, PICP reliability plots (Goovaerts, 2001) were created by calculating PICP across multiple nominal prediction interval levels. These plots were used to assess whether under- or overestimation of uncertainty occurred consistently across the predictive distribution, or only at specific interval levels. For QRF, CQR, and LU-CQR, the lower and upper bounds were based on predicted quantiles. For the other methods, lower and upper bounds were defined by symmetric conformity-score cut-offs around the point prediction. Although these bounds are not direct quantiles of a predictive distribution, they define prediction intervals at the selected nominal coverage level and were therefore evaluated using PICP.

2.3.2 Prediction interval width (PIW) was used to assess the sharpness of the predicted uncertainty. For each observation, width was calculated as the difference between the upper and lower interval bounds. While PICP describes whether the intervals contain the expected proportion of observations, it does not show how wide the intervals are. Therefore, interval width was used to compare how much uncertainty each method expressed for the same nominal coverage level, with particular attention given to PI90.

2.3.3 Interval Score (IS) was used to combine reliability and sharpness into one diagnostic. Schmidinger and Heuvelink (2023) proposed IS as a useful scoring rule for comparing probabilistic predictions in DSM. In contrast to PICP and similar metrics, which only record whether an observation falls inside or outside the interval, IS also accounts for how far outside the interval an uncovered observation lies. A lower score is given to narrow intervals, but a penalty is added when the observed value falls below or above the interval bounds. The penalty increases with the distance from the interval boundary, meaning that large misses are penalized more strongly than small misses.

Lower IS values therefore indicate better prediction intervals, because the intervals are both sharper and better calibrated. IS was calculated for each observation and then summarized using the mean interval score for each method and interval level.

2.3.4 Quantile Coverage Probability (QCP) and Probability Integral Transform (PIT). Because PICP does not show whether the lower and upper interval bounds are correctly balanced, QCP and PIT diagnostics were calculated for the baseline QRF model, following recommendations from Schmidinger and Heuvelink (2023). These diagnostics were used to detect possible one-sided bias in the predictive distribution, where overall interval coverage may appear adequate but observations are unevenly distributed around the predicted quantiles.

2.4 Area of applicability and extrapolation risk

Because prediction intervals may not fully capture uncertainty caused by extrapolation to poorly represented covariate combinations, the Area of Applicability (AOA) approach of Meyer and Pebesma (2021) was used as an additional diagnostic. The AOA is based on the dissimilarity between prediction locations and the training data in predictor space. Using feature importances from the baseline QRF model, a dissimilarity index (DI) was calculated for prediction locations across Estonia. Locations with DI values below the applicability threshold were considered inside the AOA, while locations exceeding the threshold were considered outside the AOA.

The DI and AOA layers were compared with the spatial distribution of prediction interval widths from all uncertainty quantification methods. This comparison was used to assess whether areas with higher predictor-space dissimilarity also tended to have wider prediction intervals. Spearman rank correlation was calculated between DI and prediction interval width for each method, and spatial overlays were used to identify areas where high model-based uncertainty coincided with extrapolation risk.

3. Results

3.1 Comparison of uncertainty quantification methods using diagnostics

3.1.1 Baseline QRF diagnostics. The baseline QRF model produced generally well-calibrated SOC prediction intervals. Marginal PICP for the 90% prediction interval was approximately 0.92, indicating a slight overestimation of uncertainty, but overall close agreement with the nominal coverage level. Coverage varied somewhat between land use classes and predicted SOC concentration classes. Among land use classes, forest was the best calibrated, with PICP close to 0.90, while wetlands showed the strongest over-coverage, with PICP around 0.97.

The PICP and QCP reliability plots in Figure 2 show the same overall pattern. The QCP curves indicate generally good quantile-level calibration, with small land use-specific deviations. Arable land, and to a lesser extent forest, show a slight tendency for QCP to lie above the ideal line, suggesting that predicted quantiles may be shifted a bit upward relative to the observations. Wetlands show a rather unstable pattern, which may partly reflect the small number of wetland observations in the validation data (n = 35). PIT histograms showed the same pattern: quantiles predicted for arable land were slightly high, and wetlands showed a more irregular distribution, but not a clear one-sided bias.

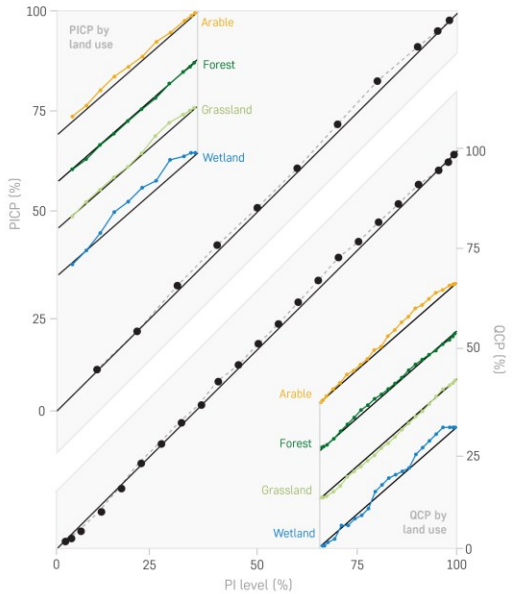

Figure 2. Combined PICP and QCP reliability plot for the baseline QRF model. Black points show marginal reliability across all validation observations. Coloured lines show land use-specific deviations. Black lines indicate ideal calibration. Deviations above or below the diagonal indicate over- or under-coverage at the corresponding PI level.

3.1.2 Comparison of uncertainty quantification methods. Tables 1 and 2 summarize the performance of all tested uncertainty quantification methods in terms of coverage and sharpness, while Figure 3 provides a visual comparison using PICP reliability plots and interval scores.

Coverage. In terms of marginal coverage, CQR achieved the closest agreement with the nominal PI90 level, with PICP equal to 0.90. RFoCS showed the strongest over-coverage, with PICP around 0.94. Overall, all methods produced reasonably good marginal coverage, with slight deviations toward overestimation rather than underestimation of uncertainty.

The land use-specific diagnostics showed more variation between methods. Arable land and grassland tended to show slight over-coverage for most methods, while forest was generally close to nominal coverage. Wetlands showed the largest differences between methods, likely partly because they were the least represented land use class in the dataset. Both QRF and CQR overestimated uncertainty for wetlands, while LU-CQR achieved coverage closer to the nominal level, with wetland PICP around 0.89. This was also the land use class where the difference between CQR and LU-CQR was most visible, suggesting that class-specific conformal calibration had the clearest effect for wetlands.

Method	PI90 width, mean					PI90 width, median				
	Total	Arable	Forest	Grassland	Wetland	Total	Arable	Forest	Grassland	Wetland
QRF	19.69*	13.31*	21.02*	19.98	38.00	15.63*	11.22	14.33*	19.87	19.18*
CQR	20.42	13.95	21.98	20.16	40.34	15.70	6.96*	17.57	16.55*	43.97
LU-CQR	20.58	14.18	22.52	21.00	29.67	16.43	7.05	18.15	17.45	31.97
RFoCS	21.28	17.32	23.01	18.53*	42.59	18.05	13.49	19.63	17.12	40.84
LU-CP	27.77	32.32	26.42	27.07	24.07*	25.87	34.92	25.87	24.73	26.45

Table 1. Mean and median width of 90% prediction intervals, by method and land use class. Lowest values marked with *

Method	PICP				
	Total	Arable	Forest	Grassland	Wetland
QRF	0.92	0.93	0.90	0.94	0.97
CQR	0.90*	0.89*	0.88	0.92	1
LU-CQR	0.92	0.94	0.90	0.93	0.89
RFoCS	0.94	0.95	0.94	0.92	0.94
LU-CP	0.91	0.93	0.90*	0.92*	0.91*

Table 2. PICP for 90% prediction intervals, by method and land use class. Values closest to the nominal 0.90 level are marked with *.

However, the PICP reliability plots show that nominal PI90 coverage alone does not fully describe interval performance. For example, LU-CP achieved close to nominal coverage across land use classes, but its interval scores and average interval widths were less favourable than those of several other methods.

In wetlands, LU-CQR and LU-CP achieved the closest PI90 coverage, but their reliability curves indicate stronger over-coverage at lower PI levels and slight under-coverage at the highest PI levels. In contrast, QRF and CQR showed more consistent over-coverage across several nominal PI levels. Therefore, while LU-CQR and LU-CP performed well at the PI90 level for wetlands, QRF and CQR showed more stable calibration across the predictive distribution.

When grouped by predicted SOC concentration, the methods differed more clearly. RFoCS and LU-CP tended to overestimate uncertainty at low predicted SOC values. This is likely related to their symmetric interval structure around the point prediction, which can be problematic near the lower boundary of SOC values. At intermediate to high SOC concentrations, the same methods showed stronger under-coverage, while LU-CP behaved irregularly in the highest SOC class. Results suggest that interval shape and whether the method allows asymmetric uncertainty are important when SOC values are close to zero or highly skewed.

Sharpness. Figure 4 shows the average PI90 width by method and land use class, together with the average asymmetry of the intervals where applicable. For QRF, CQR, and LU-CQR, prediction intervals for arable land, forest, and grassland were strongly asymmetric, with substantially wider upper than lower bounds. On average, the upper side of the interval was roughly three times wider than the lower side for these land use classes. This pattern was not true for wetlands, where intervals were wider but symmetric.

Wetlands had the widest intervals overall for most methods, consistent with their higher SOC variability and smaller sample representation. However, LU-CP produced the narrowest wetland intervals while also achieving close to nominal PI90 coverage. This suggests that simple land use-conditional calibration captured some useful class-level differences in uncertainty, even though it does not model within-class variation.

Overall, SOC prediction intervals were very wide relative to the predicted values. When normalized by the median prediction, PI90 width was several times larger than the predicted SOC value for some land use classes. For example, average PI90 width was around four times larger than the median prediction for arable land, while the relative width was lower for wetlands, at approximately 1.5 times the predicted value. Although absolute interval width increased with predicted SOC, the interval width normalized by the predicted value generally decreased as SOC concentration increased.

Interval Score. Interval scores were very similar for QRF, CQR, and LU-CQR, with LU-CQR having the lowest mean interval score, although only by a small margin. RFoCS ranked fourth overall and remained relatively close to the QRF-based methods. LU-CP had the highest interval score, indicating that although it performed well in terms of coverage, this was achieved with less favourable sharpness – as was expected for such a robust method. Land use- and SOC concentration-specific interval scores are shown in Figure 3.

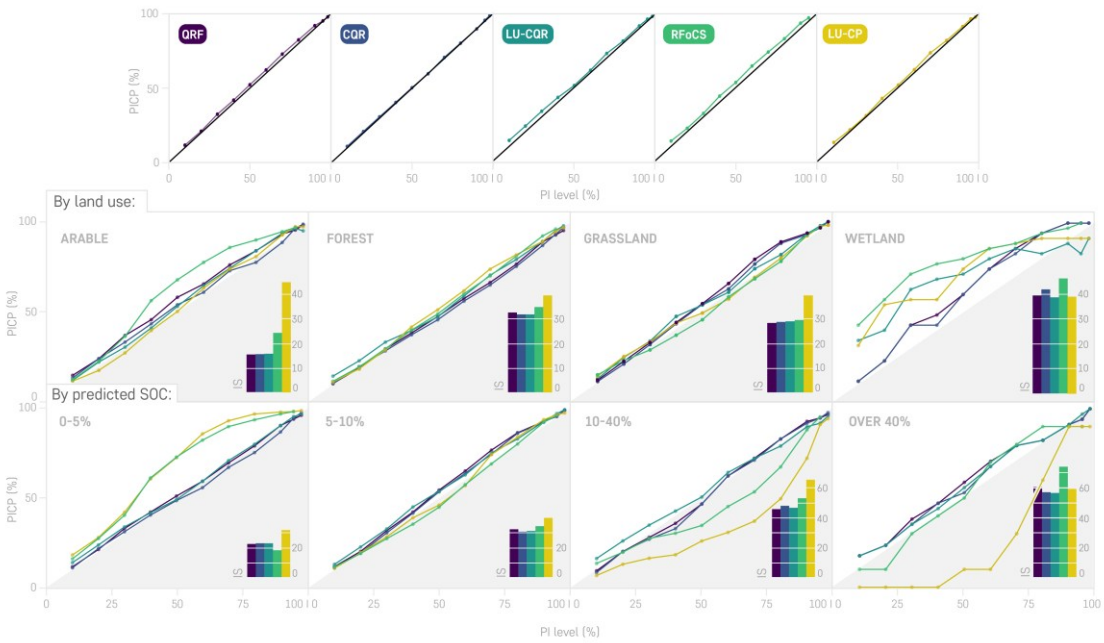

Figure 3. PICP reliability plots and interval scores for all five methods. The top row shows marginal PICP reliability across all validation observations. The middle and lower panels show PICP reliability by land use class and predicted SOC class, respectively. Insets show mean interval scores, where lower values indicate better combined reliability and sharpness.

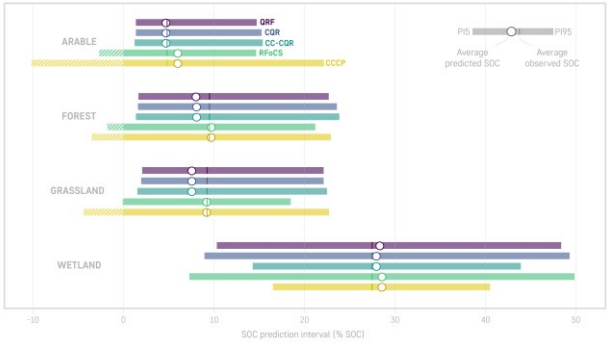

Figure 4. Average PI90 width and interval asymmetry by method and land use class.

3.2 Environmental patterns in prediction interval width

To explore how high-uncertainty predictions were distributed across the training data and environmental covariates, Figure 5 visualizes all training observations ordered by predicted SOC and coloured by PI width. This provided a visual data mining approach for comparing where different methods assigned high uncertainty and whether these patterns corresponded to particular covariate values.

Several patterns are visible in Figure 5. First, within the low to moderate SOC range, especially around predicted SOC values of approximately 4–6%, there is a cluster where uncertainty increases across all methods. This region appears to coincide with somewhat higher clay fraction, but the specific reason for this increase would be worthy of further analysis. Second, the figure makes visible which observations contributed to the land use-specific PICP and interval width patterns. For example, some wetland observations received noticeably narrower intervals under LU-CQR than under the other QRF-based methods.

Most high-uncertainty patterns were shared between methods, but RFoCS differed in some parts of the SOC gradient. From approximately 20% SOC upward, RFoCS did not show the same relationship with clay fraction as the other methods. In QRF, CQR, and LU-CQR, intervals tended to be narrower where clay content was lower, while this pattern was less evident for RFoCS.

Figure 6 provides a more systematic comparison between high uncertainty and environmental covariates for four of the five tested methods, stratified by predicted SOC concentration. Overall, the relationships between high uncertainty and environmental covariates were broadly similar across methods, but the strength and direction of some patterns varied by SOC range.

Several covariates changed their relationship with uncertainty across the predicted SOC gradient. For example, high values of topographic variables such as slope and TRI were less prominent in high-uncertainty areas at low SOC concentrations, but became more prominent in high-uncertainty areas at high SOC concentrations. TWI showed the opposite pattern. Similar shifts were visible for several soil and vegetation-related covariates. These results indicate that high uncertainty was not associated with single covariates in a uniform way, but depended on the combination of environmental conditions and predicted SOC level.

The largest method-specific differences were visible for NDVI, BSI, and the dissimilarity index. For NDVI in summer, QRF and the CQR-based methods showed lower uncertainty in high-NDVI, high-SOC areas, while low-SOC areas with high NDVI showed somewhat higher uncertainty. RFoCS showed a different pattern, with higher summer NDVI linked to slightly higher uncertainty at high SOC values. For BSI, high values were strongly associated with high uncertainty at medium-high SOC levels for the CQR-based methods, while the same pattern was less evident for the other methods.

3.3. Area of Applicability and extrapolation risk

Across Estonia, only around 1.7% of pixels were classified as outside the Area of Applicability. This indicates that the SOC model had generally very good coverage of the predictor space, with only a small share of prediction locations falling outside the range of conditions represented by the training data.

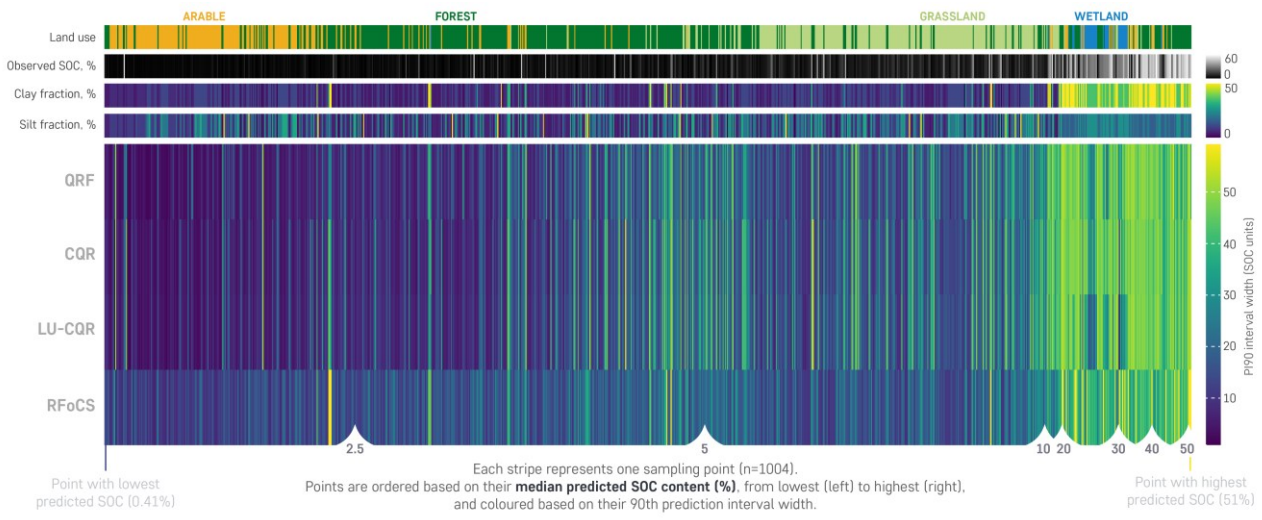

Figure 5. Observation-level comparison of PI90 width across uncertainty quantification methods. Each vertical stripe represents one sampling point, ordered from lowest to highest median predicted SOC, along with point-level covariate data. LU-CP is not shown because it produces one interval width per land use class, making it structurally different from observation-level methods shown here.

https://doi.org/10.5194/isprs-archives-L-4-W1-2026-63-2026

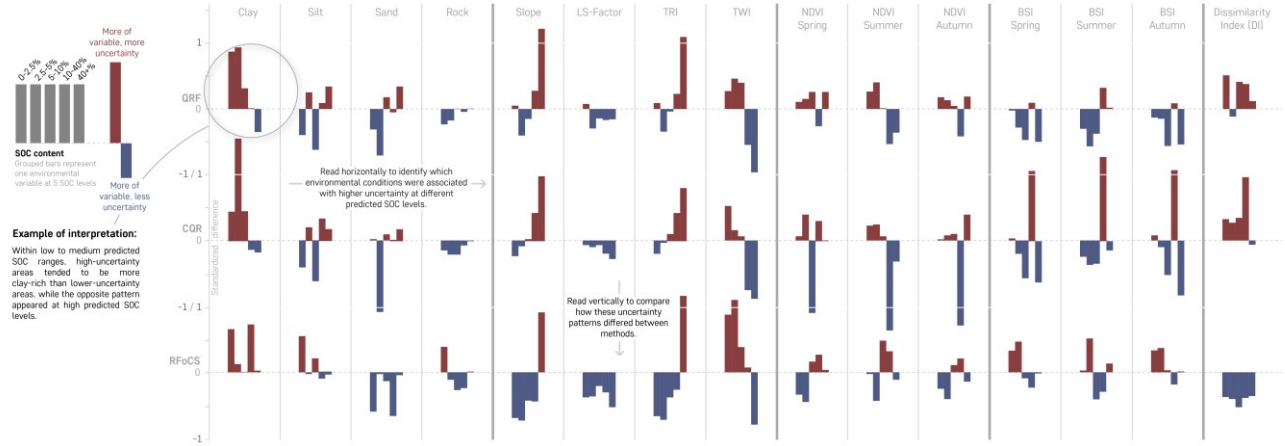

Figure 6. Environmental covariate patterns associated with high PI90 width. High uncertainty was defined as the upper quartile within each method and predicted SOC class. Bars show standardized mean differences in covariate values between high- and lower-uncertainty predictions. LU-CQR is omitted because its patterns were similar to CQR, and LU-CP because its class-level intervals are not directly comparable.

The relationship between predictor-space dissimilarity and prediction interval width was weak. Spearman correlations between DI and PI90 width were very low for most methods. RFoCS was the exception, showing a somewhat stronger – and negative – relationship, with $\rho = -0.24$. In general, locations outside the AOA had slightly wider prediction intervals than locations inside the AOA, but this pattern was not consistent across all methods or regions.

Some areas with high DI or outside-AOA conditions did not necessarily have very wide prediction intervals, while wide prediction intervals also occurred inside the AOA. This indicates that prediction intervals alone do not fully communicate whether the model is extrapolating to poorly represented covariate combinations. Conversely, AOA and DI do not replace interval-based uncertainty diagnostics, because high model uncertainty can occur even in areas that are well represented in predictor space.

4. Discussion and conclusions

4.1 Comparison of methods

The comparison of interval-based uncertainty quantification methods showed that all tested methods captured broadly similar uncertainty patterns, but also that each method contributed somewhat different information. QRF performed well as a general baseline for SOC uncertainty prediction, with good marginal calibration and interpretable asymmetric prediction intervals. Applying CQR to QRF improved marginal coverage, but did not substantially improve the sharpness of the intervals in most categories. This is likely partly because the baseline QRF intervals were already close to nominal coverage.

The land use-conditional version of CQR was most useful for wetlands, where QRF and standard CQR tended to overestimate uncertainty. This suggests that even a simple group-conditional calibration can help when prediction errors differ between broad land use classes. However, the wetland class also had the smallest number of observations, so the stability of this correction has to be interpreted with caution. LU-CP was the most conservative method overall and had the highest interval score, but it still provided useful comparative information because it showed how much uncertainty each land use class received when only class-level residual distributions were used. The opposite behaviour of LU-CP compared with the other methods – especially arable land receiving the widest intervals and wetlands the narrowest – should be investigated further.

RFoCS showed somewhat different behaviour from the QRF-based methods, especially in relation to some environmental covariates. This indicates that modelling conformity scores directly may capture uncertainty patterns that are not expressed in the same way by quantile-based methods. Future work should look more closely at what drives the local cut-off values produced by RFoCS, and whether these patterns reflect meaningful local error structure.

4.2 Environmental patterns in uncertainty

High uncertainty was linked to specific covariate combinations rather than to individual covariates in a uniform way. For example, high slope and TRI values were less common in high-uncertainty areas at low SOC concentrations, but became clearly more prominent in high-uncertainty areas at high SOC concentrations. TWI showed the opposite pattern. This likely reflects the model's lower confidence in less common environmental combinations. Since high slope generally coincides less often with high SOC, locations where both occur together represent a more unusual combination for the model. Similarly, high TWI is commonly associated with wetter and more organic-rich conditions, so low-SOC locations with high TWI may also fall into a less familiar part of the covariate space. The same logic could be applied to texture – clay-rich soils are often favourable for higher SOC storage, so low-SOC areas with high clay content are less common, and thus more difficult for the model to predict confidently.

This shows that prediction interval width could theoretically help identify not only where uncertainty is high, but also which environmental combinations may be underrepresented in the training data. This would make uncertainty patterns useful for future sampling design: collecting additional samples from these less common covariate combinations could directly target areas where the model currently has the least confidence.

4.3 Area of Applicability

The weak correlation between DI/AOA and SOC prediction interval width indicates that these diagnostics describe different

uncertainty problems. High DI areas did not consistently have wider intervals, while some high-uncertainty areas occurred inside the AOA. Together with the environmental covariate analysis, this suggests that prediction interval patterns can provide additional information for future sampling design, alongside AOA-based approaches such as Choi et al. (2025). Even inside the AOA, some covariate combinations appear to remain difficult for the model, and additional samples from these conditions could help reduce future uncertainty. Conversely, a narrow prediction interval does not necessarily mean that the model is not extrapolating. For this reason, PI-based uncertainty estimates should ideally be reported together with AOA or similar diagnostics, or predictions outside the AOA excluded.

4.4 Evaluating uncertainty predictions

Overall, the results support the use of multiple diagnostics when evaluating uncertainty predictions. PI90 PICP is useful and easy to interpret, but it is not sufficient on its own. Reliability plots, QCP, PIT, interval width, and interval score each showed different aspects of method behaviour. Breaking the diagnostics down by land use and predicted SOC class was especially important, because good marginal coverage did not always mean equally good performance within subgroups. In future work, it may be useful not only to compare uncertainty quantification methods, but also to consider whether several uncertainty expressions could be reported together, since different methods appeared to perform better under different environmental conditions.

Data and code availability

The full Python workflow used in this study will be made available after publication at https://github.com/liinah/soc-uncertainty-quantification.

Acknowledgements

This work was funded by the Estonian Research Agency (grant numbers PRG1764, PSG841, PRG1604), Estonian Ministry of Education and Research (Centre of Excellence for Sustainable Land Use - TK232, Estonian Centre of Excellence in Artificial Intelligence - TK213), the European Commission (ERC, WaterSmartLand, 101125476). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

Arrouays, D., Grundy, M. G., Hartemink, A. E., Hempel, J. W., Heuvelink, G. B., Hong, S. Y., ... & Zhang, G. L., 2014. GlobalSoilMap: Toward a fine-resolution global grid of soil properties. *Advances in Agronomy*, 125, 93-134. doi.org/10.1016/B978-0-12-800137-0.00003-0.

Cabezas, L. M., Otto, M. P., Izbicki, R., & Stern, R. B., 2025. Regression trees for fast and adaptive prediction intervals. *Information Sciences*, 686, 121369. doi.org/10.1016/j.ins.2024.121369.

Choi, J., Kmoch, A., & Uuemaa, E., 2025. Optimisation of sampling design for multivariate soil mapping with machine learning. *Proceedings of the 2025 Conference on Big Data from Space (BiDS'25)*, 201-204. doi.org/10.2760/2119408.

Goovaerts, P., 2001. Geostatistical modelling of uncertainty in soil science. *Geoderma*, 103(1-2), 3-26. doi.org/10.1016/S0016-7061(01)00067-2.

Helm, A., 2023. Role of Grasslands in Mitigating Climate Change, 2022-2023. EIC Environmental Program Project. University of Tartu.

Kasraei, B., Heung, B., Saurette, D. D., Schmidt, M. G., Bulmer, C. E., & Bethel, W., 2021. Quantile regression as a generic approach for estimating uncertainty of digital soil maps produced from machine-learning. *Environmental Modelling & Software*, 144, 105139. doi.org/10.1016/j.envsoft.2021.105139.

Kmoch, A., Kanal, A., Astover, A., Kull, A., Virro, H., Helm, A., Pärtel, M., Ostonen, I., & Uuemaa, E., 2021. EstSoil-EH: A high-resolution eco-hydrological modelling parameters dataset for Estonia. *Earth System Science Data*, 13(1), 83-97. doi.org/10.5194/essd-13-83-2021.

Meinshausen, N., 2006. Quantile regression forests. *Journal of Machine Learning Research*, 7(6).

Meyer, H., & Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, 12(9), 1620-1633. doi.org/10.1111/2041-210X.13650.

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., & Fernández-Ugalde, O., 2018. LUCAS Soil, the largest expandable soil dataset for Europe: A review. *European Journal of Soil Science*, 69(1), 140-153. doi.org/10.1111/ejss.12499.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85), 2825-2830.

Poggio, L., De Sousa, L. M., Batjes, N. H., Heuvelink, G., 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. doi.org/10.5194/soil-7-217-2021.

Romano, Y., Patterson, E., & Candes, E., 2019. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32. doi.org/10.48550/arXiv.1905.03222.

Schmidinger, J., & Heuvelink, G. B., 2023. Validation of uncertainty predictions in digital soil mapping. *Geoderma*, 437, 116585. doi.org/10.1016/j.geoderma.2023.116585.

Shrestha, D. L., & Solomatine, D. P., 2006. Machine learning approaches for estimation of prediction interval for the model output. *Neural Networks*, 19(2), 225-235. doi.org/10.1016/j.neunet.2006.01.012.

Zhang, H., Zimmerman, J., Nettleton, D., & Nordman, D. J., 2020. Random forest prediction intervals. *The American Statistician*. doi.org/10.1080/00031305.2019.1585288.