https://doi.org/10.5194/egusphere-2024-3703
Preprint. Discussion started: 7 January 2025

Using Monte Carlo conformal prediction to evaluate the uncertainty of deep learning soil spectral models

Yin-Chung Huang, José Padarian, Budiman Minasny, and Alex B. McBratney
School of Life and Environmental Science & Sydney Institute of Agriculture, The University of Sydney, NSW, Australia
Corresponding to: Yin-Chung Huang (lloyd.yc.huang@sydney.edu.au)

Abstract. Uncertainty quantification is a crucial step for the practical application of soil spectral models, particularly in supporting real-world decision making and risk assessment. While machine learning has made remarkable strides in predicting various physiochemical properties of soils using spectroscopy, predictions devoid of quantified uncertainty offer limited utility in guiding critical decisions. However, uncertainty quantification remains underutilised in the reporting of soil spectral models, with existing methods facing significant limitations. These approaches are either computationally demanding, fail to achieve the desired coverage of observed data, or struggle to handle out-of-domain uncertainty effectively. This study introduces the innovative use of Monte Carlo conformal prediction (MC-CP) as a novel approach to quantify uncertainty in the prediction of clay content from mid-infrared spectroscopy. We compared MC-CP with two established methods: (1) Monte Carlo dropout and (2) conformal prediction. Monte Carlo dropout generates prediction intervals for each sample and is effective at addressing larger uncertainties associated with out-of-domain data. However, it falls short in achieving the desired coverage – its 90% prediction intervals only covered the observed values in 74% of cases, well below the expected 90% coverage. Conformal prediction, on the other hand, guarantees ideal coverage of true values but generates unnecessarily wide prediction intervals, making it overly conservative for many practical applications. In contrast, MC-CP successfully combines the strengths of both methods. It achieved a prediction interval coverage probability of 91%, closely matching the expected 90% coverage, and far surpassing the performance of Monte Carlo dropout. Additionally, the mean prediction interval width for MC-CP was 9.05%, narrower than conformal prediction's 11.11%, while still effectively addressing the higher uncertainty in out-of-domain samples. By generating accurate prediction intervals alongside point predictions, MC-CP demonstrated its ability to deliver practical and reliable uncertainty quantification. This breakthrough enhances the real-world applicability of soil spectral models and represents a significant advancement in the field of soil science. The success of MC-CP paves the way for its integration into large-scale machine-learning models, such as soil inference systems, further revolutionising decision-making and risk assessment in soil science.

1 Introduction

In the recent developments of soil science, machine learning has been widely used, such as soil spectroscopy, proximal sensing, carbon stock modelling, and digital soil mapping (Padarian et al., 2020; Minasny et al., 2024). These studies are characterised by the use of large soil datasets and require an efficient way of extracting information to predict target attributes. Hence, machine learning is favoured because these algorithms can generate prediction models with high accuracy for various purposes. For example, in soil spectroscopy, visible and near-infrared (Vis-NIR) and mid-infrared (MIR) spectroscopy have been used with machine learning to predict soil properties such as soil organic carbon (SOC), texture, and cation exchange capacity (CEC). Padarian et al. (2019) applied convolutional neural networks (CNN) to predict soil properties with Vis-NIR spectra, using the European Land Use/Cover Area Frame Statistical Survey (LUCAS) dataset which contained about 20,000 samples. In the SOC prediction of their study, the multi-task CNN outperformed conventional algorithms, such as partial least squares regression and Cubist, by reducing the root mean square error (RMSE) by more than 60%. Additionally, Ng et al. (2019) used

https://doi.org/10.5194/egusphere-2024-3703
Preprint. Discussion started: 7 January 2025

the Kellogg Soil Survey Laboratory (KSSL) database with around 15,000 samples from the United States (US) with both Vis-NIR and MIR spectra to build a multi-task CNN. Their model achieved a coefficient of determination (R^2) over 0.90 for total carbon, SOC, CEC, clay, sand, and pH.

Despite the significant success of machine learning in predicting soil properties, uncertainty quantification of the prediction remained an underexplored area in soil spectroscopy, and only a few studies have tried to include uncertainty in the model evaluation. Uncertainty quantification is a crucial factor to consider when dealing with models in soil spectroscopy, as quantifying the confidence of models in predicting soil properties is equally important to providing the predictions themselves. The growing demand for practical applications of soil spectral models requires users to know the uncertainty accompanying the model prediction to assess the quality of the predictions (Bellon-Maurel et al., 2010). In the data-intensive context of deep learning (DL), uncertainty analysis is critical in evaluating models for decision-making and risk management, and predictions without uncertainty are neither practicable nor applicable (Begoli et al., 2019). Hence, it is crucial to establish an effective way to evaluate the uncertainty of machine learning models.

Two types of uncertainty are generally considered in machine learning, namely aleatoric and epistemic uncertainty (Hüllermeier and Waegeman, 2021). Aleatoric uncertainty is inherent in the randomness of the experiments and is not reducible by additional sampling or knowledge. In contrast, epistemic uncertainty is caused by the lack of information, or in other words, "by ignorance", and can be improved by increasing the amount of information involved in the model (Hüllermeier and Waegeman, 2021). Epistemic uncertainty is the main topic in this study.

An ideal uncertainty quantification method is expected to satisfy the following criteria:

(1) the method is computationally efficient,
(2) the prediction interval coverage probability (PICP) must meet the expected coverage, i.e. a p% coverage is expected for a p% prediction interval, with the narrowest mean prediction interval width (MPIW), and
(3) the prediction intervals should be able to address the larger uncertainty for samples significantly different from training set (i.e. out-of-domain samples).

Several methods have been used to generate prediction intervals for each prediction to characterise epistemic uncertainty. One commonly used approach is bootstrapping, in which several models are trained with subsets generated by drawing samples with replacements from the same dataset (Efron and Tibshirani, 1994). The mean of all models is considered the final prediction, and a prediction interval can be derived from the quantiles of multiple predictions. However, one drawback of bootstrapping is the time-consuming nature of training numerous bootstrapping models, in addition, bootstrapping only deals with the model uncertainty. A comprehensive uncertainty quantification using methods such as Markov Chain Monte Carlo can better evaluate the parameter uncertainty involved in the model (Minasny et al., 2011).

An alternative method to evaluate model uncertainty is the Monte Carlo dropout (MC dropout) by Gal and Ghahramani (2016), in which a CNN model is trained with multiple dropout layers that randomly deactivate neurons during prediction, resulting in different predictions across iterations. Multiple predictions from a single MC dropout CNN model form a distribution, and prediction intervals could be obtained by assessing the quantiles of the predictions. This approach reduced the training time compared to bootstrapping.

The performance of bootstrapping and MC dropout was compared by Padarian et al. (2022), in which CNN models were trained to predict SOC with Vis-NIR spectra using the LUCAS dataset through (1) a hundred times bootstrapping and (2) MC dropout. Additionally, CNN models were trained on mineral soils with a threshold of <20% SOC and then separately tested on in-domain data (mineral soils, SOC<20%) and out-of-domain data (organic soils, SOC>20%). This was to test the model's response to samples significantly different from the training set. A good uncertainty quantification should indicate the larger uncertainty when predicting out-of-domain data. When facing in-domain data, both bootstrapping and MC dropout generated reasonable prediction intervals. However, when facing out-of-domain data, the prediction interval of MC dropout increased

https://doi.org/10.5194/egusphere-2024-3703
Preprint. Discussion started: 7 January 2025

significantly compared to bootstrapping, indicating that the uncertainty increased when the testing samples were markedly different from the training data. In other words, the model was aware of its uncertainty for out-of-domain data and can reflect this situation by generating a wider prediction interval. Such analysis is particularly useful when assessing risk management, as predictions with higher uncertainty must be treated cautiously. However, both bootstrapping and MC dropout underestimated the uncertainty and were over-confident in their study. The 90% PICP of bootstrapping and MC dropout in their study were both under 80% while the expected coverage was 90%. This was not practical in real-world situations and left room for improvement.

A relatively easier method to generate prediction intervals with expected coverage is conformal prediction (CP), which uses an independent calibration set to estimate the prediction interval and can be performed on any model (Shafer and Vovk, 2008). Kakhani et al. (2024) utilised CP to generate prediction intervals for SOC mapping in Europe with the LUCAS dataset, and CP outperformed every other method by generating the most accurate PICP and a reasonably-sized prediction interval. Despite these advantages, a key limitation of CP is its inability to generate sample-specific prediction intervals. Instead, it produces a uniform interval for all samples. In other words, CP does not account for increased uncertainty in out-of-domain samples. As a result, CP is known as a conservative method that provides overly broad prediction intervals. Consequently, no uncertainty quantification method applied in soil spectroscopy has yet combined computational efficiency, expected coverage with a narrow MPIW, and the ability to address out-of-domain uncertainty.

In this study, we applied a strategy to increase the PICP of MC dropout while maintaining its advantages in characterising out-of-domain uncertainty. Monte Carlo-Conformal Prediction (MC-CP) was introduced by Bethell et al. (2024). MC-CP integrates the strengths of both MC dropout and CP. It not only retains the structure of the MC dropout to generate different prediction intervals for each sample but also extends the prediction interval with CP to achieve the expected coverage. In other words, MC-CP can guarantee expected coverage while accounting for the uncertainty of each sample. Bethell et al. (2024) demonstrated the effectiveness of MC-CP in both regression and classification tasks using benchmark datasets and showed that MC-CP was significantly improved from the original MC dropout. Hence, MC-CP is a promising method for soil science and can address the uncertainty involved in prediction.

This study aimed to explore the use of MC-CP as a potential method to quantify the uncertainty of DL models in soil spectroscopy. Specifically, the goal was to validate whether MC-CP preserves the advantages of both MC dropout and CP. Therefore, the objectives of this study are to (1) test if MC-CP can generate prediction intervals that reach expected PICP, and (2) evaluate if MC-CP can address the uncertainty of out-of-domain samples.

2 Materials and methods

2.1 Dataset

The soil samples from the Kellogg Soil Survey Laboratory (KSSL) dataset were used in this study. It contained the MIR spectra and physiochemical properties of over 17,000 soil profiles and 70,000 soil samples across the US (Soil Survey Staff, 2014). Soil clay content was selected as the target variable to predict with MIR in this study, as the prediction of clay has been a well-established method for MIR spectroscopy (Seybold et al., 2019; Ng et al., 2022). Since the spectra behave differently in mineral and organic soils, the samples with SOC>10% were removed. Extreme values for clay content were also removed by excluding data below the 5th percentile and above the 95th percentile, leaving a total number of 39,177 samples.

Here we created a model based on the in-domain data, and a threshold of 40% clay content was chosen to separate the in-domain and out-of-domain samples. This resulted in approximately 10% number of samples being classified as out-of-domain (clay>40%, N=3,686) compared to the in-domain samples (clay<40%, N=35,491). The in-domain samples would be used for

https://doi.org/10.5194/egusphere-2024-3703
Preprint. Discussion started: 7 January 2025

model training, validation, and testing. The out-of-domain samples were not involved in any of the training processes and were only used to test the performance of models when facing out-of-domain situations.

The clay contents were scaled to a range of 0-1 using the maximum and minimum of the training set. The spectra were trimmed to 4000-600 cm^{-1} for analysis, and the full procedure of MIR spectral analysis can be found in the manual by Soil Survey Staff (2014). No other preprocessing was applied to the raw spectra as it has been proved that CNNs are able to deal with spectra without preprocessing (Ng et al., 2019; Padarian et al., 2019).

2.2 Monte Carlo dropout

MC dropout was introduced by Gal and Ghahramani (2016) based on dropout layers, which are commonly used in DL models to prevent overfitting (Srivastava et al., 2014). In each dropout layer, a certain portion of the neurons is randomly deactivated (weights set to zero) during both training and testing. By randomly dropping neurons and their connections, the dropout layer helps prevent the model from adapting too much to the training dataset. As a result of the dropout layers, each prediction result is different, and multiple predictions generate a distribution. Gal and Ghahramani (2016) demonstrated that MC dropout can be used to approximate Bayesian inference in deep Gaussian processes, and the standard deviation of the prediction can thus be used for assessing the uncertainty (Bethell et al., 2024). For a detailed rationale, readers are referred to the original paper by Gal and Ghahramani (2016).

In practice, a CNN model with dropout layers was trained to predict 100 times to generate a distribution. The 90% prediction interval of each sample i would be calculated as the difference between the 5th quantile ($\hat{q}_5(X_i)$) and the 95th quantile ($\hat{q}_{95}(X_i)$) of the predictions (Eq. 1).

$$C_{MC}(X_i) = [\hat{q}_5(X_i),\ \hat{q}_{95}(X_i)] \qquad \text{Eq. 1}$$

2.3 Conformal prediction (CP)

CP is a model-agnostic method, which means it can be used to evaluate the uncertainty of any model (Shafer and Vovk, 2008). Consider (X_i, Y_i), $\mathrm{i} = 1,2,\ldots \mathrm{n}$ to be pairs of features (inputs) and responses (outputs), and α is the desired error level. A regression model f is constructed using the training dataset, and $f(X_i)$ is the prediction of the observed value Y_i. The goal is to generate prediction intervals $C(X_i)$ such that the probability of the observed value Y_i being contained within $C(X_i)$ is approximately $1-\alpha$ (Angelopoulos and Bates, 2022). The procedure can be separated into three steps:

(1) **Start with nonconformity scores.** The nonconformity measure is the foundation of CP, which quantifies how different the predicted values are from the observed values (Shafer and Vovk, 2008). In a regression scenario, the nonconformity measure is typically defined as the absolute value of residuals $z_i = |f(X_i) - Y_i|$. Here, z_i represents the nonconformity scores of the i-th data point. The first step is to calculate these nonconformity scores using a calibration dataset and rank the nonconformity scores from low to high. Table 1 shows an example dataset of 100 samples with the z_i in the order from minimum to maximum.

https://doi.org/10.5194/egusphere-2024-3703
Preprint. Discussion started: 7 January 2025

Table 1: Example dataset of conformal prediction containing 100 samples. The nonconformity scores are ranked from minimum to maximum.

N	$f(X)$	Y	$z = \lvert f(X) - Y \rvert$ (Nonconformity scores)	$C(X_i) = [f(X_i) - \hat{q}, f(X_i) + \hat{q}]$
1	96	95.9	0.1	[93.8, 98.2]
2	3	3.2	0.2	[0.8, 5.2]
3	96	95.7	0.3	[93.8, 98.2]
4	18	18.4	0.4	[15.8, 20.2]
5	71	70.5	0.5	[68.8, 73.2]
6	99	99.6	0.6	[96.8, 101.2]
7	38	37.3	0.7	[35.8, 40.2]
8	11	11.8	0.8	[8.8, 13.2]
9	74	73.1	0.9	[71.8, 76.2]
10	54	55	1.0	[51.8, 56.2]
			…	
91	24	21.9	2.1	[21.8, 26.2]
92	56	58.2	**2.2**	[53.8, 58.2]
93	48	45.5	2.5	[45.8, 50.2]
94	19	21.8	2.8	[16.8, 21.2]
95	90	86.9	3.1	[87.8, 92.2]
96	27	30.2	3.2	[24.8, 29.2]
97	70	66.6	3.4	[67.8, 72.2]
98	66	69.9	3.9	[63.8, 68.2]
99	21	16.8	4.2	[18.8, 23.2]
100	80	84.5	4.5	[77.8, 82.2]

https://doi.org/10.5194/egusphere-2024-3703
Preprint. Discussion started: 7 January 2025

(2) **Get adjusted quantile.** Using the ranked nonconformity scores, CP computes the adjusted quantile to determine the prediction interval. Specifically, it selects the $\frac{\lceil (1-\alpha)(n+1) \rceil}{n}$-th quantile of the nonconformity scores z_i to be $\hat{q}$. The $\lceil\ \rceil$ symbol indicates the ceiling function, and this equation is to correct the quantile for the size of the calibration dataset (Angelopoulos and Bates, 2022). In the example, if we set $\alpha = 0.1$ with a total of 100 samples, the $\hat{q}$ will be the 92nd quantile of z_i, which is 2.2 (marked as bold in Table 1).

(3) **Generate prediction interval.** The prediction intervals are constructed as Eq. 2.

$$C_{CP}(X_i) = [f(X_i) - \hat{q},\ \ f(X_i) + \hat{q}] \qquad \text{Eq. 2}$$

The width of each prediction interval is fixed to two times the value of $\hat{q}$, centred around the model prediction $f(X_i)$. In the example from Table 1, the prediction interval covered the observed values from sample 1 through 92, indicating 92% of the samples are covered in the prediction interval. This $\hat{q}$ will be applied to testing set to generate prediction intervals for unknown data. The key advantage of CP is that it can be applied to any model, regardless of correctness, assumptions, or structure of the model while providing guaranteed coverage for the specified confidence level (Angelopoulos and Bates, 2022). However, the fixed interval width for all data points and guaranteed coverage also makes CP an over-conservative method that generates unnecessarily wide intervals (Bethell et al., 2024).

2.4 Monte Carlo-conformal prediction (MC-CP)

MC-CP is a novel uncertainty quantification method developed by Bethell et al. (2024). As its name suggests, MC-CP combined MC and CP to estimate the uncertainty. Instead of using CP to generate a prediction interval, MC-CP extends the prediction interval from an MC method. While deep quantile regression was used in the original paper by Bethell et al. (2024) to generate prediction intervals, a CNN with dropout layers was used in the present study to give point estimates. The CNN model with dropout layers is trained in the same way as the MC dropout method to predict the calibration set 100 times. For each sample i in the calibration set the 5th quantile ($\hat{q}_5(X_i)$) and the 95th quantile ($\hat{q}_{95}(X_i)$) of the 100 predictions are calculated, and the nonconformity score E_i is defined as Eq. 3.

$$E_i := Max\{\hat{q}_5(X_i) - Y_i,\ \ Y_i - \hat{q}_{95}(X_i)\} \qquad \text{Eq. 3}$$

According to Eq. 3, the nonconformity scores are calculated as the largest distance between the observed value and the boundary of the MC dropout interval. The $\frac{\lceil (1-\alpha)(n+1) \rceil}{n}$-th quantile of the nonconformity scores E_i will then be selected to be $\hat{Q}$. The adjusted prediction interval of MC-CP will be calculated as Eq. 4.

$$C_{MC-CP}(X_i) = \left[\hat{q}_5(X_i) - \hat{Q},\ \ \hat{q}_{95}(X_i) + \hat{Q}\right] \qquad \text{Eq. 4}$$

In MC-CP, the prediction interval of the MC dropout method will be extended by two times $\hat{Q}$. For unknown testing data, the prediction interval will first be calculated in the same way as the MC method and then extended by two times $\hat{Q}$, which is calculated from the calibration set. This will result in sample-dependent prediction intervals, guaranteed coverage, and less conservative intervals than CP.

https://doi.org/10.5194/egusphere-2024-3703
Preprint. Discussion started: 7 January 2025

2.5 Model architecture and training data

A 1D CNN was constructed with five trainable layers, namely four convolutional layers and one fully-connected (dense) layer. The detailed description of layers is shown in Table 2. A fixed filter size of five was used for all convolutional layers, and the filter size for the max-pooling layer was fixed at two. The number of filters started at 32 and increased to 256. Every convolutional layer was followed by a max-pooling layer and an MC dropout layer, so four dropout layers were applied with a fixed 20% dropout rate. Dropout rates including 10%, 20%, and 30% were tested and optimised. The network was trained with a batch size of 300, a maximum epoch of 500, and early stopping on the validation set with patience set as 60. The initial learning rate was set to 0.001, and the learning rate reduction factor was set to 0.1 with the patience set as 50. These hyperparameters were also tested and optimised for this dataset.

Table 2: Architecture of the convolutional neural network. ReLU stands for rectified linear unit.

Layer type	Filter size	Filters	Activation
Convolutional	5	32	ReLU
Max-Pooling	2		
MC Dropout (0.2)			
Convolutional	5	64	ReLU
Max-pooling	2		
MC Dropout (0.2)			
Convolutional	5	128	ReLU
Max-pooling	2		
MC Dropout (0.2)			
Convolutional	5	256	ReLU
Max-pooling	2		
MC Dropout (0.2)			
Flatten			
Fully-connected			Linear

The in-domain data was separated into 85% training, 5% validation, 5% CP calibration, and 5% testing. Only the training and validation data were used in building the CNN model. All the analyses were performed in Python v3.12.3 using Tensorflow v2.16.1 (Abadi and Zheng, 2015; Python Software Foundation, 2024).

2.6 Model evaluation

The model performance was evaluated using coefficients of determination (R^2, Eq. 5) and root mean squared error (RMSE, Eq. 6).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad \text{Eq. 5}$$

https://doi.org/10.5194/egusphere-2024-3703
Preprint. Discussion started: 7 January 2025

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad \text{Eq. 6}$$

The results of uncertainty quantification were evaluated using the prediction interval coverage probability (PICP, Eq. 7) and the mean prediction interval width (MPIW, Eq. 8) following (Shrestha and Solomatine, 2006):

$$PICP = \frac{1}{n} count\ j \quad \text{Eq. 7}$$
$$j: PL_i^L \le y_i \le PL_i^U$$

$$MPIW = \frac{1}{n}\sum_{i=1}^{n}[PL_i^U - PL_i^L] \quad \text{Eq. 8}$$

where n is the total number of observations, and j is the number of samples in which the observed value y_i is covered in the prediction interval. PL_i^L and PL_i^U are the lower and upper bounds of the prediction interval of the i-th sample. PICP calculates the proportion that the true value is covered by the interval, while MPIW calculates the average length of prediction intervals.

3 Results and discussion

3.1 Model performance

The DL model demonstrated a good performance in predicting the clay content of in-domain test set, with R^2 of 0.90 and RMSE of 3.39% (Table 3). The results were comparable to those of the multi-task CNN models by Ng et al. (2019), which used part of the current dataset. For out-of-domain samples, the performance mostly deviated from the 1:1 line and had poor R^2 (Fig. 1; Table 3). Such results for out-of-domain samples were expected as the model did not have any knowledge of soils with clay content larger than 40%. As a result, most of the out-of-domain predictions fell under 40% clay.

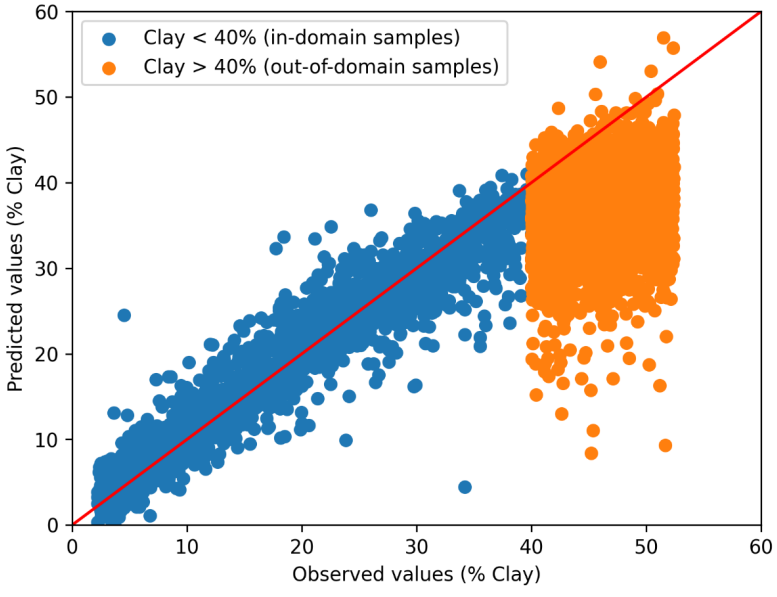

Figure 1: Relationship between the observed and predicted clay content (%) of the convolutional neural network model for in-domain testing set and out-of-domain samples.

https://doi.org/10.5194/egusphere-2024-3703
Preprint. Discussion started: 7 January 2025

Table 3: Results of the convolutional neural network modelling. R^2 stands for coefficient of determination, and RMSE stands for root mean square error.

	In-domain test set (n=1775)	Out-of-domain (n=3686)
R^2	0.90	-6.64
RMSE (%)	3.39	9.65

3.2 Uncertainty quantification

Uncertainty quantification served as an evaluation for the prediction intervals. When a model was predicting with higher uncertainty (in the case of out-of-domain samples), the models are expected to generate wider MPIW to indicate its lack of knowledge. Padarian et al. (2022) demonstrated that MC dropout possessed the ability to "know what they know" and produced prediction intervals for out-of-domain samples five times larger than in-domain samples.

Prediction intervals were generated by making 100 predictions of each sample (Fig. 2), and PICP refers to the probability that this interval covers the observed value. When the evaluation of uncertainty is optimal, the expected coverage of a p% prediction interval is p% (dotted line in Fig. 3) (Shrestha and Solomatine, 2006). In the current study, the MC dropout continuously underestimated the uncertainty through all prediction intervals (Fig. 3). This trend was similar to the finding of Padarian et al. (2022). In contrast, the PICP of CP and MC-CP were both close to the expected coverage (Fig. 3). This is attributed to the "guaranteed coverage" features of CP, and MC-CP provides an augmented effect.

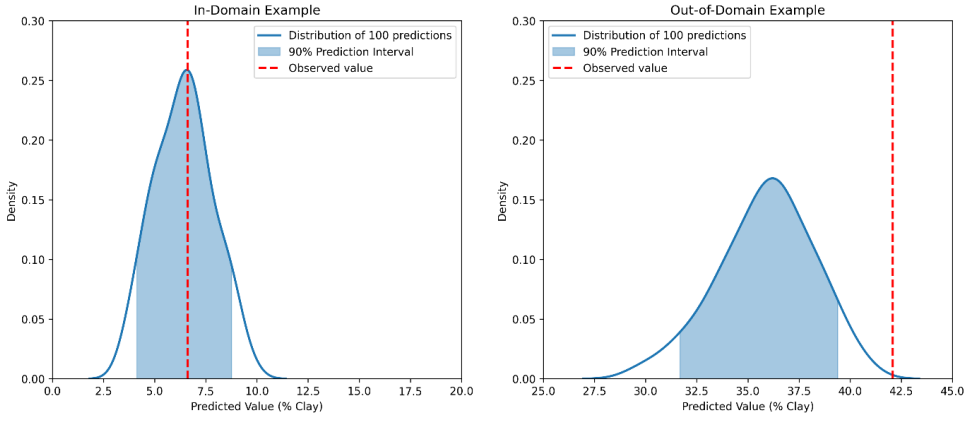

Figure 2: Examples of the distribution of 100 predictions for an in-domain and an out-of-domain sample using MC dropout. Shaded areas are the 90% prediction interval. The 90% prediction interval of the in-domain example covered the observed value while the 90% prediction interval of the out-of-domain example did not cover the observed value.

https://doi.org/10.5194/egusphere-2024-3703
Preprint. Discussion started: 7 January 2025

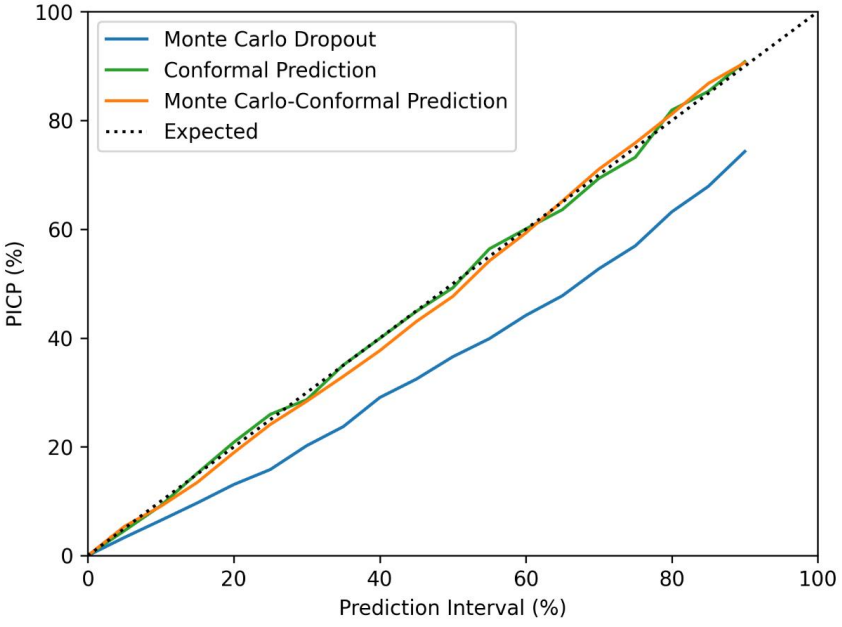

Figure 3: Prediction interval coverage probability (PICP) of in-domain samples at different prediction intervals for Monte Carlo dropout, conformal prediction, and Monte Carlo-conformal prediction.

For 90% prediction intervals, the MC dropout method achieved 74% coverage (Fig. 3; Table 4), indicating an overconfident interval. The MPIW of MC dropout for in-domain testing samples was 5.56%, the narrowest of all three methods (Table 4). This is further supported by the distribution of PIW in Fig. 4. When encountering out-of-domain samples, the MPIW of MC dropout was 6.93%, 25% higher than the MPIW of in-domain samples. This demonstrated the ability of MC dropout to generate wider intervals when encountering samples they are not familiar with (Padarian et al., 2022).

Table 4: Results of uncertainty quantification by Monte Carlo dropout, conformal prediction, and Monte Carlo-conformal prediction. PICP stands for prediction interval coverage probability, and MPIW stands for mean prediction interval width.

Method	90% PICP in-domain	MPIW (%) in-domain	MPIW (%) out-of-domain
Monte Carlo dropout	74%	5.56	6.93
Conformal prediction	91%	11.11	11.11
Monte Carlo-conformal prediction	91%	9.05	10.42

https://doi.org/10.5194/egusphere-2024-3703
Preprint. Discussion started: 7 January 2025

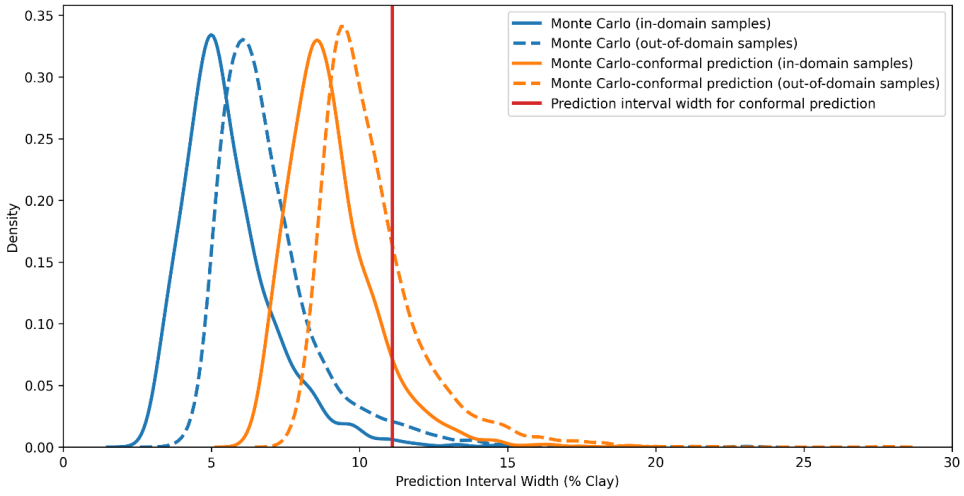

Figure 4: Distribution of prediction interval width of Monte Carlo dropout, conformal prediction, and Monte Carlo-conformal prediction for in-domain and out-of-domain samples.

On the other hand, both CP and MC-CP were able to achieve a coverage of 91% (Fig. 3; Table 4), from the expected coverage of 90%. This implied that 91% of the prediction interval contained the true observed clay content, making the prediction interval reliable. However, the MPIWs of CP and MC-CP were higher than the MPIW of MC, indicating a trade-off between narrower interval and coverage. The MPIW of CP (11.11%) was the largest of the three methods, twice of MC dropout (5.56%) (Fig. 4; Table 4), making it overly conservative. Additionally, the interval of CP was constant, this prohibited CP from addressing the different uncertainties as "guaranteed coverage" was the main objective of this method. In other words, CP generated wide prediction intervals that were unnecessary.

The MC-CP method achieved a balance between MC dropout and CP, which produced an MPIW between MC dropout and CP while still reaching the expected coverage. The MPIW of MC-CP was 9.05%, which was 1.6 times the MPIW of MC (Table 4) but achieved a 91% coverage from the expected coverage of 90%. Additionally, MC-CP retained the ability to address the uncertainty of out-of-domain samples, as the MPIW for out-of-domain samples (10.42%) was larger than the MPIW for in-domain samples (9.05%) (Fig. 4; Table 4). Hence, MC-CP is an adequate compromise among (1) coverage of observed values, (2) addressing out-of-domain uncertainty, and (3) reasonably sized MPIW.

3.3 Limitations and future applications

The MC-CP method was able to quantify the uncertainty and generated prediction intervals with sufficient coverage of true values. However, one obvious difference between MC-CP and MC is that MC-CP requires calibration samples to establish nonconformity scores. However, only a small number of calibration samples were needed compared to the training samples, and this can easily be done by dividing a portion of the training sample. For instance, the size of the calibration set in the MC-CP regression example presented by Bethell et al. (2024) was only 2% of the testing samples. Future studies could explore determining the optimal size of calibration sets.

While CP is model-agnostic, MC dropout is restricted to deep neural networks since it requires the inclusion of dropout layers in the model architecture (Gal and Ghahramani, 2016). Hence, MC-CP is also model-specific and can only be used on deep neural networks. Neural networks have been widely applied in soil spectroscopy, with several studies reporting accurate prediction results (Ng et al., 2019; Padarian et al., 2019; Javadi et al., 2021). By including dropout layers during training and

https://doi.org/10.5194/egusphere-2024-3703
Preprint. Discussion started: 7 January 2025

separating calibration samples, MC-CP can be applied across various scenarios. Additionally, MC-CP can also be considered an alternative when the MC dropout result is not ideal since it is a stand-alone method and does not take much effort to upgrade to MC-CP. Comparing MC-CP with methods that generate prediction intervals differently could also be valuable. For example, Bayesian CNNs incorporate uncertainty in the model itself and produce probabilistic predictions. In soil spectroscopy, Bayesian CNNs were found to outperform bootstrapped PLS and achieve the expected PICP (Omondiagbe et al., 2024). Future investigations should compare these uncertainty quantification methods concerning computational efficiency and MPIW.

In this study, the PICP for out-of-domain samples was low, with only 26% coverage for the MC-CP method. This was because the CNN model lacked information about the out-of-domain samples. It would be worth testing if the inclusion of out-of-domain samples into CP calibration can improve the coverage. That is, to spike samples in CP calibration instead of training samples. The "guaranteed coverage" nature of CP can theoretically achieve the same coverage, but it might produce an extremely wide prediction interval to compensate for the low coverage by MC dropout. Additionally, Schmidinger and Heuvelink (2023) raised the issue that PICP ignores the one-sided bias in prediction, in which 90% of the interval covers the observed value but the probability outside the boundaries is asymmetrically distributed. Other parameters are thus needed to evaluate the uncertainty quantification in the future.

MC-CP is expected to be applied to large models such as soil inference systems (McBratney et al., 2002), in which multiple pedotransfer functions were coupled together to predict complicated soil properties using basic soil properties that can be assessed from soil spectroscopy. Adding uncertainty analysis into model evaluation will increase the practicality of models and bring them one step closer to real-world applications.

4 Conclusions

The study aimed to assess the uncertainty in predicting clay content using a deep learning (DL) method through three uncertainty quantification techniques: Monte Carlo (MC) dropout, conformal prediction (CP), and MC-CP. The mid-infrared (MIR) spectra from the KSSL database were divided into two categories:

- In-domain samples
- Out-of-domain samples: This division tested the model's ability to handle samples that differ significantly from the training data.

The following methods were compared:

- MC Dropout:
  - Produced the lowest Prediction Interval Coverage Probability (PICP).
  - Generated the narrowest Mean Prediction Interval Width (MPIW), indicating overconfidence in predictions.
- Conformal Prediction (CP):
  - Achieved the ideal PICP but had a fixed and the largest MPIW among the methods.
- MC-CP:
  - Balanced the strengths of the other methods, achieving 91% PICP (for a 90% expected PICP) with a moderate MPIW.

The advantages of MC-CP are:

- Provides an optimal trade-off in uncertainty quantification.
- Exhibits:
  1. High coverage probability of true values.
  2. Variable prediction intervals that adapt to out-of-domain samples.

https://doi.org/10.5194/egusphere-2024-3703
Preprint. Discussion started: 7 January 2025

3. Moderate MPIW for balanced uncertainty representation.

The main implications are:

- MC-CP demonstrates the potential for quantifying uncertainty in DL models for soil property prediction.
- The method allows for computationally efficient uncertainty quantification, producing prediction intervals that reliably cover the true values and address out-of-domain uncertainties.

Future directions:

- Integration of MC-CP into large-scale prediction systems, such as soil inference models, to enhance prediction accuracy and support decision-making in real-world applications.

Acknowledgements

The authors acknowledge the staff at the National Soil Survey Center Kellogg Soil Survey Laboratory (Lincoln, NE) who have collected and analysed the soil samples in the KSSL dataset. We would like to thank the authors of Bethell et al. (2024), especially Daniel Bethell, for providing the original code of MC-CP. The study acknowledges funding from National Soil Carbon Innovation Challenge – Development and Demonstration Round 2 grant: An integrated schema for soil carbon stock estimation and crediting.

Data availability

The data used in this study are owned and managed by the National Soil Survey Center Kellogg Soil Survey Laboratory (KSSL). Interested parties should contact KSSL directly to request access, in accordance with their data-sharing policies.

Author contributions

Yin-Chung Huang: Conceptualization, Data curation, Formal analysis, Investigation, Visualization, Writing – original draft preparation; José Padarian: Conceptualization, Supervision, Writing – review & editing; Budiman Minasny: Supervision, Writing – review & editing; Alex B. McBratney: Supervision, Writing – review & editing.

Competing interests

The authors declare that they have no conflict of interest.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X.: TensorFlow: large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/, 2015.

Angelopoulos, A. N., and Bates, S.: A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification, arXiv, 2107.07511, https://doi.org/10.48550/arXiv.2107.07511, 2022.

Begoli, E., Bhattacharya, T., and Kusnezov, D.: The need for uncertainty quantification in machine-assisted medical decision making, Nat. Mach. Intell., 1(1), 20-23, https://doi.org/10.1038/s42256-018-0004-1, 2019.

https://doi.org/10.5194/egusphere-2024-3703
Preprint. Discussion started: 7 January 2025

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.-M., and McBratney, A.: Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy, TrAC, Trends Anal. Chem., 29(9), 1073-1081, https://doi.org/10.1016/j.trac.2010.05.006, 2010.

Bethell, D., Gerasimou, S., and Calinescu, R.: Robust Uncertainty Quantification Using Conformalised Monte Carlo Prediction, Proceedings of the AAAI Conference on Artificial Intelligence, 38(19), 20939-20948, https://doi.org/10.1609/aaai.v38i19.30084, 2024.

Efron, B., and Tibshirani, R. J.: An Introduction to the Bootstrap. Chapman and Hall/CRC, New York, NY. https://doi.org/10.1201/9780429246593, 1994.

Gal, Y., and Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, Proceedings of The 33rd International Conference on Machine Learning, 48, 1050-1059, https://proceedings.mlr.press/v48/gal16.html, 2016.

Hüllermeier, E., and Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods, Mach. Learn., 110(3), 457-506, https://doi.org/10.1007/s10994-021-05946-3, 2021.

Javadi, S. H., Munnaf, M. A., and Mouazen, A. M.: Fusion of Vis-NIR and XRF spectra for estimation of key soil attributes, Geoderma, 385, 114851, https://doi.org/10.1016/j.geoderma.2020.114851, 2021.

Kakhani, N., Alamdar, S., Kebonye, N. M., Amani, M., and Scholten, T.: Uncertainty Quantification of Soil Organic Carbon Estimation from Remote Sensing Data with Conformal Prediction, Remote Sens., 16(3), 438, https://doi.org/10.3390/rs16030438, 2024.

McBratney, A. B., Minasny, B., Cattle, S. R., and Vervoort, R. W.: From pedotransfer functions to soil inference systems, Geoderma, 109(1), 41-73, https://doi.org/10.1016/S0016-7061(02)00139-8, 2002.

Minasny, B., Bandai, T., Ghezzehei, T. A., Huang, Y.-C., Ma, Y., McBratney, A. B., Ng, W., Norouzi, S., Padarian, J., Rudiyanto, Sharififar, A., Styc, Q., and Widyastuti, M.: Soil Science-Informed Machine Learning, Geoderma, 452, 117094, https://doi.org/10.1016/j.geoderma.2024.117094, 2024.

Minasny, B., Vrugt, J. A., and McBratney, A. B.: Confronting uncertainty in model-based geostatistics using Markov Chain Monte Carlo simulation, Geoderma, 163(3), 150-162, https://doi.org/10.1016/j.geoderma.2011.03.011, 2011.

Ng, W., Minasny, B., Jeon, S. H., and McBratney, A.: Mid-infrared spectroscopy for accurate measurement of an extensive set of soil properties for assessing soil functions, Soil Secur., 6, 100043, https://doi.org/10.1016/j.soisec.2022.100043, 2022.

Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., and McBratney, A. B.: Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra, Geoderma, 352, 251-267, https://doi.org/10.1016/j.geoderma.2019.06.016, 2019.

Omondiagbe, O. P., Roudier, P., Lilburne, L., Ma, Y., and McNeill, S.: Quantifying uncertainty in the prediction of soil properties using mid-infrared spectra, Geoderma, 448, 116954, https://doi.org/10.1016/j.geoderma.2024.116954, 2024.

Padarian, J., Minasny, B., and McBratney, A. B.: Using deep learning to predict soil properties from regional spectral data, Geoderma Reg., 16, e00198, https://doi.org/10.1016/j.geodrs.2018.e00198, 2019.

Padarian, J., Minasny, B., and McBratney, A. B.: Machine learning and soil sciences: a review aided by machine learning tools, Soil, 6(1), 35-52, https://doi.org/10.5194/soil-6-35-2020, 2020.

Padarian, J., Minasny, B., and McBratney, A. B.: Assessing the uncertainty of deep learning soil spectral models using Monte Carlo dropout, Geoderma, 425, 116063, https://doi.org/10.1016/j.geoderma.2022.116063, 2022.

Python Software Foundation: Python Language Reference, version 3.12.3. https://www.python.org, 2024.

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

https://doi.org/10.5194/egusphere-2024-3703
Preprint. Discussion started: 7 January 2025

Seybold, C. A., Ferguson, R., Wysocki, D., Bailey, S., Anderson, J., Nester, B., Schoeneberger, P., Wills, S., Libohova, Z., Hoover, D., and Thomas, P.: Application of Mid-Infrared Spectroscopy in Soil Survey, Soil Sci. Soc. Am. J., 83(6), 1746-1759, https://doi.org/10.2136/sssaj2019.06.0205, 2019.

Shafer, G., and Vovk, V.: A tutorial on conformal prediction, J. Mach. Learn. Res., 9, https://doi.org/10.48550/arXiv.0706.3188, 2008.

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

Soil Survey Staff: Kellogg soil survey laboratory methods manual, Soil Survey Investigations Report No. 42. N. R. C. S. United States Department of Agriculture, 2014.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, J. Mach. Learn. Res., 15, 1929-1958, 2014.