Referee #2

General Revision Summary

The study presents a well-executed analysis of uncertainty quantification in deep learning soil spectral models, specifically through the use of Monte Carlo-Conformal Prediction (MC-CP). The paper makes a strong contribution to the field by addressing a crucial gap in soil spectroscopy—reliable uncertainty quantification. The methodological approach is well-documented, and the comparison between MC dropout, Conformal Prediction (CP), and MC-CP is insightful and thorough.

Reply: Thank you for your positive comments.

Strengths of the Paper

- Novel Contribution: The paper introduces MC-CP as an method for uncertainty quantification in deep learning soil spectral models. The demonstration of MC-CP's ability to balance expected coverage, computational efficiency, and adaptability to out-of-domain samples is a significant advancement.
- Well-Designed Comparison: The comparison between MC dropout, CP, and MC-CP is informative and shows the trade-offs between these methods.
- Strong Methodological Foundation: The study follows a solid methodological framework.
- Practical Relevance: The application of the proposed method to real-world soil spectral data enhances the practical impact of the study

Reply: Thank you for your positive comments.

General Improvements

Terminology and Consistency (Machine Learning vs. Deep Learning)

- The abstract and introduction interchangeably refer to Machine Learning (ML) and Deep Learning (DL). However, the methodology and model used are specifically deep learning-based. Ensure consistency in terminology and explicitly state where ML is a broader category and where DL is specifically applied.

Reply: Thank you. We added text to state that ML is the broader category, and the specific method used in this study is for deep learning.

"Additionally, deep learning (DL), as a branch of machine learning, is increasingly being applied to soil science to explore its ability to extract information from large datasets."

"This study introduces an innovative application of Monte Carlo conformal prediction (MC-CP) to quantify uncertainty in deep learning models for predicting clay content from mid-infrared spectroscopy."

"An alternative method to evaluate model uncertainty in DL is the Monte Carlo dropout (MC dropout) by Gal and Ghahramani (2016)"

"This study aimed to explore the use of MC-CP as a potential method to quantify the uncertainty of DL models in soil spectroscopy."

- Incorporate a broader range of examples in the Introduction for Monte Carlo (MC) Dropout and Conformal Prediction (CP), as the current section focuses too narrowly on just two detailed examples.

Reply: We added an example by Singh et al. (2024) using conformal prediction for earth observation data.

"Singh et al. (2024) applied CP with ML in earth observation data, and CP successfully generated prediction intervals of canopy height."

Reference:

Singh, G., Moncrieff, G., Venter, Z., Cawse-Nicholson, K., Slingsby, J., and Robinson, T. B.:

Uncertainty quantification for probabilistic machine learning in earth observation using conformal prediction, Scientific Reports, 14(1), 16166, https://doi.org/10.1038/s41598-024-65954-w, 2024.

- There is a lack of clear structure (detailed further in the comments). There are many redundant repetitions, and subordinate clauses with very general information are interspersed throughout, often repeating details that were already mentioned earlier.

Reply: We address the issues in the detailed comments section.

- Ensure there is a space between the number and the percentage symbol for proper formatting.

Reply: We corrected it accordingly.

Detailed Comments

Comment 1; L7-10

“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.”

The sentence effectively explains that predictions without uncertainty are not useful for decision-making and that uncertainty quantification is rarely used due to limitations. However, the logical connection between these points could be clearer to improve readability and coherence.

Reply: Thank you. We revised the sentences to improve readability and coherence.

“While machine learning has made remarkable strides in predicting various physiochemical properties of soils using spectroscopy, its practical utility in decision-making remains limited without quantified uncertainty. Despite its importance, uncertainty quantification is rarely incorporated into soil spectral models, with existing methods facing significant limitations.”

Comment 2; L10

” These approaches are either computationally demanding….“

It is not entirely clear whether this refers to the existing methods mentioned in the previous sentence or to something else, as methods and approaches are not necessarily the same.

Reply: We revised the sentence as follows:

“Existing methods are either computationally demanding, fail to achieve the desired coverage of observed data, or struggle to handle out-of-domain uncertainty.”

Comment 3; L11-23

The structure is confusing in the sense that your method is mentioned without prior explanation, followed by the introduction of two established methods for comparison. Additionally, while introducing these methods, you already include some results. To improve clarity, consider restructuring the section by clearly separating the description of methods, the comparison, and then presenting the results.

Reply: Thank you for the suggestions. We re-organise the section to separate the methods and results.

“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 can address larger uncertainties associated with out-of-domain data. 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. Using 39,177 samples from the mid-infrared spectral library of the Kellogg Soil Survey Laboratory to build convolutional neural networks, we found that Monte Carlo dropout itself 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. 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 the Monte Carlo dropout."

Comment 4; L24-26

"This breakthrough enhances the real-world applicability of soil spectral models and represents a significant advancement in the field of soil science. [….] further revolutionising decision-making and risk assessment in soil science."

Shorten this section to two sentences, as the usefulness is stated twice. Avoid redundant explanations to improve clarity.

Reply: We revised the sentences as follows:

"The success of MC-CP enhances the real-world applicability of soil spectral models, paving the way for their integration into large-scale machine-learning models, such as soil inference systems, and further transforming decision-making and risk assessment in soil science."

Comment 5; L29

"[...](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."

The reference is incorrect, as these studies do not discuss what you describe in the following sentence.

Reply: We changed the references to Ng et al. (2019) and Wadoux et al. (2020) as they discussed the applications of machine learning in (1) soil spectroscopy using a large spectral dataset and (2) digital soil mapping with various data sizes.

Reference:

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.

Wadoux, A. M. J. C., Minasny, B., and McBratney, A. B.: Machine learning for digital soil mapping: Applications, challenges and suggested solutions, Earth-Sci. Rev., 210, 103359, https://doi.org/10.1016/j.earscirev.2020.103359, 2020.

Comment 6; L41-46

There are repetitions in the sentences without adding new content. Shorten them for conciseness.

Reply: We shorten the sentences as follows:

“Despite the significant success of machine learning in predicting soil properties, uncertainty quantification of the prediction remained an underexplored area in soil spectroscopy (Omondiagbe et al., 2024). 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).”

Comment 7; L43

“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."

A reference is needed for the studies mentioned.

Reply: We added reference Omondiagbe et al. (2024).

Reference:

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.

Comment 8; L50-54

I don't see the relevance of explaining the difference between the two types of uncertainty here, as it does not appear to be a topic in the methods section or the discussion.

Reply: Thank you for the suggestions. We remove this part about types of uncertainty.

Comment 9; L61-66

To my knowledge, bootstrapping is typically used for confidence intervals, not for prediction intervals like MC and CP. Additionally, different methods of quantile regression and Gaussian methods are missing, which would help provide a more complete introduction.

Reply: We agree. We revise our content to address these issues.

1. We added contents to indicate that bootstrapping is only used for confidence intervals instead of prediction intervals.

   "In addition, bootstrapping primarily addresses the model uncertainty and it derives confidence intervals rather than prediction intervals (Heuvelink, 2014; Wadoux, 2019)."
2. We added introduction about quantile regression and Bayesian CNNs to provide a more complete introduction.

> “The diverse nature of models enabled the development of different methods. For example, quantile regression (QR) uses a set of regression models to estimate the quantile of target variables, and the prediction interval can later be defined by the upper and lower quantiles (Kasraei et al., 2021). Additionally, quantile regression forests (QRF) and quantile regression neural networks (QRNN) are also extensions of quantile regression that apply similar principles to generate prediction intervals (Schmidinger and Heuvelink, 2023). Heuvelink et al. (2021) utilised QRF to predict the SOC for soils in Argentina with quantified uncertainty, and the 0.05 and 0.95 quantiles were used to generate the 90 % prediction interval. On the other hand, Omondiagbe et al. (2024) compared bootstrapped PLS, generalised additive models (GAM), and Bayesian CNNs for their ability to quantify uncertainty. They found that GAM and Bayesian CNN outperformed bootstrapped PLS by having PICP close to the ideal 90% value. Moreover, the MPIW of Bayesian CNN is mostly lower than that of GAM models, suggesting a more accurate estimation of uncertainty (Omondiagbe et al., 2024). However, Bayesian neural networks are more intensive in computation compared to standard CNNs (Bethell et al., 2024; Omondiagbe et al., 2024).”

References:

Heuvelink, G. B.: Uncertainty quantification of GlobalSoilMap products, GlobalSoilMap: Basis of the global spatial soil information system, 335-340, 2014.

Heuvelink, G. B. M., Angelini, M. E., Poggio, L., Bai, Z., Batjes, N. H., van den Bosch, R., Bossio, D., Estella, S., Lehmann, J., Olmedo, G. F., and Sanderman, J.: Machine learning in space and time for modelling soil organic carbon change, Eur. J. Soil Sci., 72(4), 1607-1623, https://doi.org/10.1111/ejss.12998, 2021.

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

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.

Wadoux, A. M. J. C.: Using deep learning for multivariate mapping of soil with quantified uncertainty, Geoderma, 351, 59-70, https://doi.org/10.1016/j.geoderma.2019.05.012, 2019.

Comment 10; L68-72

Specify that MC is specifically used for deep learning to avoid ambiguity.

Reply: We added "in DL" to specify that MC is used for deep learning.

"An alternative method to evaluate model uncertainty in DL is the Monte Carlo dropout (MC dropout) by Gal and Ghahramani (2016)"

Comment 11; L96-103

"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."

Clarify that MC-CP is the strategy. Again, avoid repetition to improve clarity and conciseness.

Reply: We revised the sentences to improve clarity and conciseness.

"In this study, we applied the Monte Carlo-Conformal Prediction (MC-CP), a method introduced by Bethell et al. (2024) to improve the PICP of MC dropout while maintaining its advantages in characterising out-of-domain uncertainty."

Comment 12; L113-115

Please specify how many of the removed samples were due to SOC and how many were excluded because of extreme values.

Reply: We revised the sentences to clarify how many samples were excluded in each step.

"The database contains 45,339 samples which have measured MIR spectra and particle size analysis. Since the spectra of mineral and organic soils behave differently, samples with SOC > 10 % were excluded, resulting in the removal 1,808 samples. Additionally, extreme values for clay content were also filtered by excluding data below the 5th percentile and above the 95th percentile, further removing 4,354 samples. This resulted in a total number of 39,177 samples."

Comment 13; L116

Clarify why the threshold of 40% clay content was chosen and provide justification for this choice.

Reply: We intended to create a subset of soil samples that are spectrally different from the training data. There are two reasons for choosing the 40% threshold. (1) A clay content of 40% is the minimum required for a soil to be classified as clay in most of the soil texture classification system. (2) This threshold results in an in-domain to out-of-domain sample size ratio of 10:1, which is ideal for the analysis. We updated the context to elaborate this:

"A clay content of 40 % is the minimum threshold for a soil to be classified as "clay" according to the US Department of Agriculture soil texture classification (Soil Science Division Staff, 2017). Using this criterion, approximately 10% of the samples were categorised as out-of-domain (clay>40 %, N=3,686), while the remaining in-domain samples had clay content below 40% (N=35,491)."

Comment 14; L119

If you are already describing your training and test scheme here, also include the ratio of the splitting mentioned in L203 for consistency and completeness.

Reply: Thank you for the suggestion. We moved L203 here to indicate the data split during modelling.

“The in-domain data were further randomly separated into 85 % training, 5 % validation, 5 % calibration for conformal prediction, and 5 % testing.”

Comment 15; Chapter 2.2, 2.3, 2.4

For better structure, I suggest organizing the section as follows: 2.2 Methods, with subsections 2.2.1 Monte Carlo Dropout (MC dropout), 2.2.2 Conformal Prediction (CP), and 2.2.3 Monte Carlo-Conformal Prediction (MC-CP).

Reply: We agree and organised it accordingly as 2.2 Uncertainty quantification methods, 2.2.1 Monte Carlo dropout (MC dropout), 2.2.2 Conformal prediction (CP), and 2.2.3 Monte Carlo-conformal prediction (MC-CP).

Comment 16; L125

Missing abbreviation: MC dropout

Reply: We corrected it accordingly.

Comment 17; L128

“In each dropout layer, a certain portion of the neurons is randomly deactivated (weights set to zero) during both training and testing.”

As far as I know, and as stated in the paper by Gal and Ghahramani (2016), neurons are only deactivated during training. While validation can be involved, a specific reason is needed for doing so. Please verify what is happening in your specific use case.

Reply: In the MC dropout case, the neurons are deactivated both during training and predicting. Here we refer to the paper from Bethell et al. (2024): *"Although dropout is typically used during training, MC dropout keeps this feature active during inference and performs several forward passes to devise a prediction distribution."* That is, the neurons are deactivated during training and testing.

Comment 18; L137

Check the Mathematical notation and terminology of the journal:

https://publications.copernicus.org/for_authors/manuscript_preparation.html#math.

I recommend centering the equations for better readability. Additionally, equations should be treated as nouns within the text. So here I would change it to the following:

The 90% prediction interval [...] of the predictions (Eq. 1):

Formula. (Eq. 1)

Reply: We revised it accordingly.

Comment 19; L137

When using a formula, ensure that every abbreviation is defined either before or in the sentence following it. In this case, CMC and Xi are missing definitions.

Reply: We revised it accordingly.

"In Eq. 1, X_i represents an individual input sample. The 90 % prediction interval of MC dropout ($C_{MC,\ 90}$) of each sample i would be defined by the 5th quantile ($\hat{q}_5(X_i)$) and the 95th quantile ($\hat{q}_{95}(X_i)$) of the predictions (Eq. 1)"

Comment 20; L150

Table 1

Stay consistent in using X or Xi throughout the table to maintain clarity and uniformity.

Reply: We revised it accordingly.

N	$f(X_i)$	Y_i	$r = \|f(X_i) - Y_i\|$ (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]

Comment 21; L161

See comment No. 18

Reply: We revised it accordingly.

Comment 22; L170

Stay consistent in the writing of Monte Carlo-conformal prediction. Since it is based on Bethell et al. (2024), I recommend following their terminology and formatting.

Reply: Thank you, we checked the terminology in the article. We chose to use “Monte Carlo-Conformal Prediction (MC-CP)” as the name for the method throughout the manuscript, as this name was also used by Bethell et al. (2024) in their publication. The reason for choosing

“Monte Carlo-Conformal Prediction” instead of “Conformalised Monte Carlo Prediction” is because the former can easily be understood as the combination of MC and CP.

Comment 23; L179

See comment No. 18

Reply: We revised it accordingly.

Comment 24; L184

See comment No. 18

Reply: We revised it accordingly.

Comment 25; L208-209

See comment No. 18 and a reference is missing for the Eq. 5 and 6.

Reply: We revised it accordingly and added reference Ng et al. (2022).

Reference:

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.

Comment 26; L210-214

See comment No. 18 a space is missing in Eq. 7 between the fraction and "count".

Reply: We revised it accordingly and added a space. Eq. 7 is now $PICP = \frac{1}{n}\, count\ j$.

Comment 27; L223

I would rephrase it as follows, omitting the word "poor":

“A negative R-squared value indicates that the model performs worse than simply using the mean prediction."

Reply: We agree and modified the sentence accordingly.

“For out-of-domain samples, a negative R-squared value indicates that the model performs worse than simply using the mean prediction”

Comment 28; L224-225

Connect the two sentences for example as following:

“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%, leading most out-of-domain predictions to fall under 40% clay.”

Reply: Thank you. We revised the sentence.

“Such a result for out-of-domain samples was expected, as the model lacked knowledge of soils with clay content exceeding 40 %, resulting in the most out-of-domain predictions falling below 40 % clay.”

Comment 29; L238

“When the evaluation of uncertainty is optimal, the expected coverage of a p%

prediction interval is p% (dotted line in Fig. 3)”

What do you mean by "evaluation of uncertainty"? Please clarify or provide a more precise definition.

Reply: We removed the confusing part and retain the part with expected coverage.

“The expected coverage of a p % prediction interval is p %, which is indicated by the dotted line in Fig. 3”

Comment 30; L255

MPIW instead of PIW

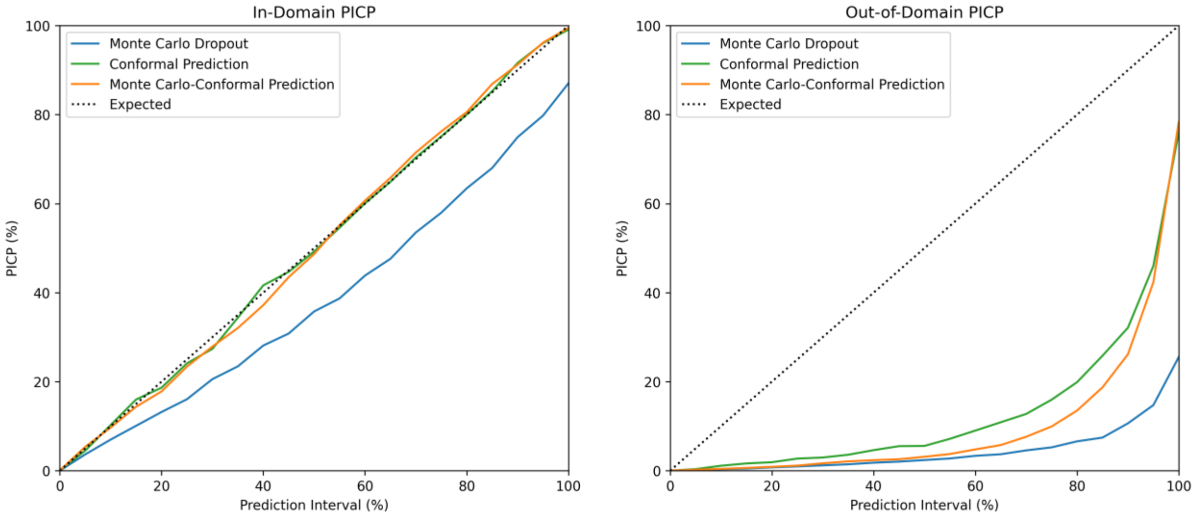
Reply: We corrected accordingly.

Comment 31; L263

Table 4

The PICP value for out-of-domain samples is missing and should be included for completeness.

Reply: We updated Fig. 3 and Table 4 to present the PICP of out-of-domain samples:



In-Domain PICP

Out-of-Domain PICP

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

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	90 % PICP out-of-domain	MPIW (%) out-of-domain

Monte Carlo dropout	74 %	5.56	11 %	6.90
Conformal prediction	91 %	11.11	32 %	11.11
Monte Carlo-conformal prediction	91 %	9.05	26 %	10.43

Comment 32; L276-281

I do not agree with the strong wording that MC-CP effectively addresses the out-of-domain issue, as the difference in MPIW between in-domain and out-of-domain samples is not significant.

Reply: Thank you for the comment. In our study, the MPIW of out-of-domain samples (10.43 %) was different from the MPIW of in-domain samples (9.05 %) by 1.38 %, this was about 15 % (1.38/9.05) increasement of width. We acknowledge that this difference is not as effective as the results in the carbon prediction in Padarian et al. (2022), which has a significant difference between the in-domain MPIW (2.73 %) and out-of-domain MPIW (15.12 %). We agree to avoid using strong words such as "effectively", and we added sentences to acknowledge that the MPIW is still not wide enough to fully address the out-of-domain samples: "However, the extended MPIW was insufficient to fully address the differences between out-of-domain samples and the in-domain training samples, and the 90 % PICP for MC dropout was only 11 %"

Additionally, we added discussion about the possible reasons behind such performance. In Liu et al. (2021), the authors found that Bayesian NN and MC dropout did not give high uncertainty to out-of-domain samples. Zadorozhny et al. (2021) discussed the problem that sometimes the NN over-generalise from training data to predict out-of-domain samples. That is, when the out-of-domain samples are similar to the in-domain samples, the algorithm wrongly assign high confidence to the prediction of out-of-domain samples. This could be the case in our study, as the spectra of clayey soils were not as distinct from those of sandy soils as the spectra of high

SOC soils were from low SOC soils. The spectra shown in Ng et al. (2022) and Zhang et al. (2022) can support this argument as the high SOC data have significant peaks at 2930-2850 cm^{-1} due to alkyl groups in SOM. Additionally, this plot only include soil samples with organic carbon less than 12%, and the difference of MIR spectra between mineral soils and organic soils could be bigger.

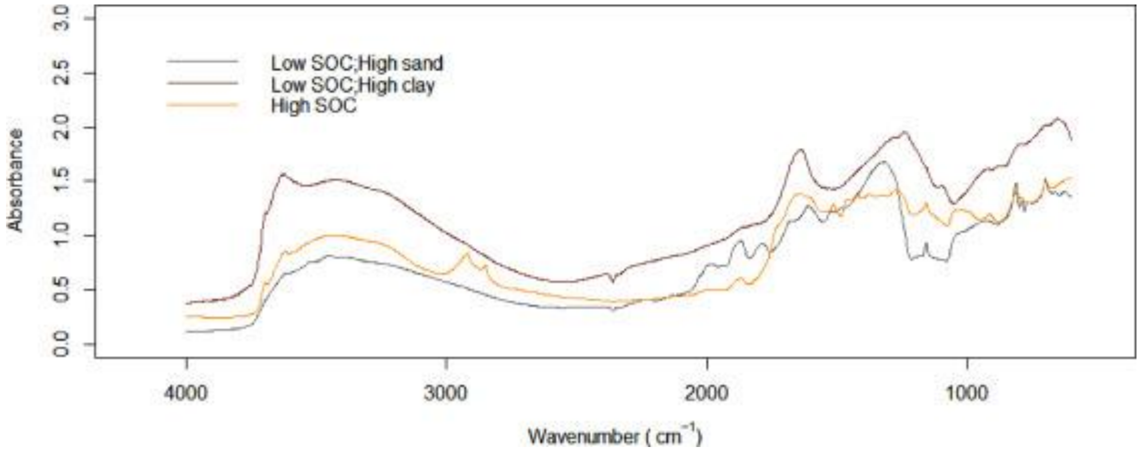


Figure from Ng et al. (2022)

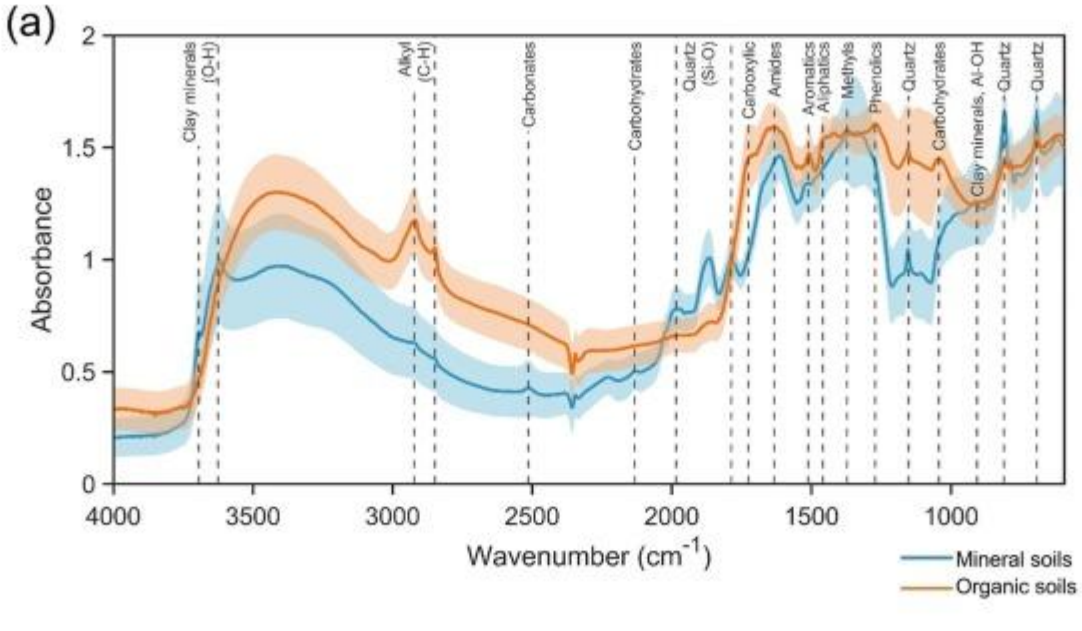


Figure from Zhang et al. (2022)

We added discussion about the possible reasons that MC-CP did not generate a wider MPIW: "However, when facing out-of-domain samples, MC-CP achieved only 26 % coverage at the 90 % prediction interval. The MPIW for out-of-domain samples (10.43 %) was 1.38 higher than that for in-domain samples (9.05 %), representing a 15 % increase in the width. The difference was insufficient to fully account for out-of-domain uncertainty, leading to the low coverage. Similarly, Liu et al. (2021) found that Bayesian neural networks and MC dropout were unable to assign high uncertainty to out-of-domain samples, indicating overconfidence in predicting unknown data. Zadorozhny et al. (2021) also highlighted the tendency of neural networks to overgeneralise from training data when predicting out-of-domain samples, potentially leading to overconfidence. When out-of-domain sample inputs closely resemble in-domain sample inputs, MC dropout may assign similar confidence levels to out-of-domain samples, failing to capture the true uncertainty. The MIR spectra of clayey soils were not as distinct from those of sandy soils as the spectra of high SOC soils were from low SOC soils (Ng et al., 2022; Zhang et al., 2022). For example, peaks at 2930-2850 cm-1 serve as a distinction between mineral soils and organic soils (Tinti et al., 2015; Ng et al., 2022). Thus, the difference between the MPIW of in-domain and out-of-domain samples was not as significant as in the study of Padarian et al. (2022), in which 20% SOC was used as the separation between in-domain and out-of-domain samples."

References:

Liu, Y., Pagliardini, M., Chavdarova, T., and Stich, S. U.: The Peril of Popular Deep Learning Uncertainty Estimation Methods, Proceedings of the Bayesian Deep Learning workshop, NeurIPS 2021, https://doi.org/10.48550/arXiv.2112.05000, 2021.

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.

Zadorozhny, K., Ulmer, D., and Cinà, G.: Failures of Uncertainty Estimation on Out-Of-Distribution Samples: Experimental Results from Medical Applications Lead to Theoretical

Insights, Proceedings of the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning, 2021.

Zhang, Y., Freedman, Z. B., Hartemink, A. E., Whitman, T., and Huang, J.: Characterizing soil microbial properties using MIR spectra across 12 ecoclimatic zones (NEON sites), Geoderma, 409, 115647, https://doi.org/10.1016/j.geoderma.2021.115647, 2022

Comment 33; L299-304

This part should be discussed directly in the uncertainty section rather than in the limitations and future applications section for better coherence.

Reply: We agree. We split the discussion about the out-of-domain samples into two parts and have an additional part in 3.2 Uncertainty quantification for better coherence. The added parts can be seen in comment 32.

Comment 34; L312

Specify the exact deep learning model used.

Reply: We specify that we used convolutional neural networks.

“The study aimed to assess the uncertainty in predicting clay content using convolutional neural networks through three uncertainty quantification techniques”

Comment 35; L329

The wording should be revised—for an optimal trade-off, the results need to be more significant.

Reply: We changed the sentence to “Provides a balance between MC Dropout and CP.”