Referee #1:

General

The authors introduce "Monte Carlo conformal prediction" (MC-CP), a combination of Monte Carlo dropout and conformal prediction, for the uncertainty quantification of soil spectroscopy predictions with deep learning models. They demonstrate its merits compared to pure Monte Carlo dropout or conformal prediction approaches for in-domain predictions. Furthermore, the paper is logically structured and easy to follow. However, the introduction describes several concepts incompletely. Moreover, I do not share their conclusions regarding the "out-of-domain" predictions, as they claim that MC-CP can address uncertainty for out-of-domain data, even though all they show is that the prediction intervals are slightly wider for out-of-domain data, without providing information on the (presumably poor) coverage. These key issues need to be addressed prior to publication. See below specific comments:

Reply: Thank you for your comments. We will address them in the following replies.

Comment 1; L. 7

The abstract should mention that Monte Carlo dropout is a method for neural networks (i.e., deep learning). Currently, it mentions “machine learning“ in L. 7, which could appear as if the method is model-agnostic for (any) machine learning model. Either, machine learning could be replaced with deep learning, or it could be somewhere else explicitly mentioned that it is a method for deep learning.

Reply: We agree. We change sentence in the abstract to mention that MC-CP is a method for deep learning.

“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.”

Comment 2; L. 12 EDIT

The more common name for this method is not "Monte Carlo conformal prediction" but "conformalised Monte Carlo prediction" (Bethell et al. 2024) following the predecessor "conformalized quantile regression" (Romano et al. 2019). However, I do not consider it a mistake because both variants exist.

Literature

Bethell, D., Gerasimou, S., & Calinescu, R. (2024). Robust uncertainty quantification using conformalised Monte Carlo prediction. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 19, pp. 20939-20948).

Romano, Y., Patterson, E., & Candes, E. (2019). Conformalized quantile regression. Advances in neural information processing systems, 32.

Reply: Thank you for your understanding. We decided to keep this name as it can indicate the combination of two methods, and this name also appears in Bethell et al. (2024).

Comment 3; L. 20 -21

I do not share the opinion that MC-CP "effectively address[ed] the higher uncertainty in out-of-domain samples" given the results presented in this paper but my discussion on that can be found in Comment 19.

Reply: Thank you. We remove the strong wording "effectively" and address this issue together in Comment 19.

Comment 4; L. 23- 26

I am convinced of the merits of MC-CP for deep learning in soil spectroscopy but the wording is exaggerated here. Neither "breakthrough" nor "revolutionizing" are appropriate terms here because the authors did not invent this method but demonstrated its advantage compared to their vanilla version for “in-domain data”.

Reply: We revise the wordings mentioned above.

“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; L. 28-29

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). Wording around “such as” sounds slightly off, and I propose adding “in applications”: “[..] widely used in applications such as soil spectroscopy, proximal sensing […]”

Reply: Thank you for the suggestion. We revised it accordingly.

“In the recent developments of soil science, machine learning has been widely used in applications such as soil spectroscopy, proximal sensing, carbon stock modelling, and digital soil mapping (Ng et al., 2019; Wadoux et al., 2020).”

Comment 6; L. 50 – 67

The concept of aleatoric and epistemic uncertainty is quite vague and may even be incorrectly applied here because of the sentence in L. 54: “Epistemic uncertainty is the main topic in this study.”

In the following, I use the definitions of Valdenegro-Toro & Mori (2022): “There are two kinds of uncertainty […]: aleatoric or data uncertainty, and epistemic or model uncertainty. These uncertainties are usually combined and predicted as a single value, called predictive uncertainty

[…].” Hence, the interest of the study is to find an uncertainty method (e.g. MC-CP) which succeeds in quantifying the combined predictive uncertainty.

Of course, epistemic uncertainty becomes especially relevant for the “out-of-domain data” predictions, as here epistemic uncertainty is very high. Hence, it is relevant that the uncertainty quantification model can also account for the epistemic uncertainty that is associated with the domain shift. Maybe this is what the authors intended to refer to but it is a far stretch to what is written and nowhere explicitly mentioned. Currently, it appears as if the authors confuse the predictive uncertainty with the epistemic uncertainty.

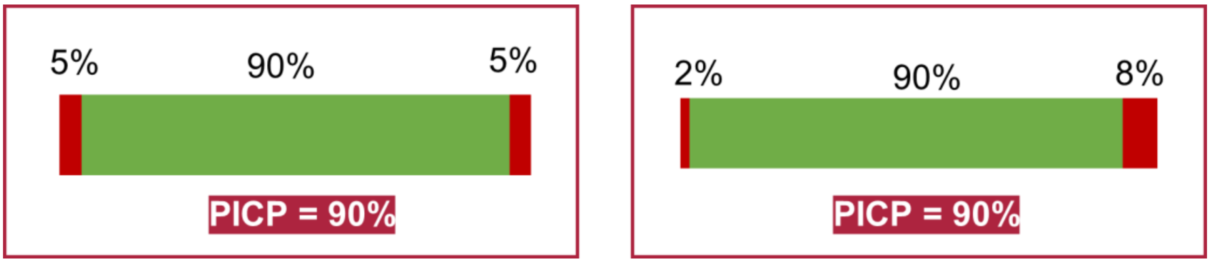
Literature

Valdenegro-Toro, M., & Mori, D. S. (2022, June). A deeper look into aleatoric and epistemic uncertainty disentanglement. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 1508-1516). IEEE.

Reply: Thank you for the suggestions. We understand that the interest of this study is on the combined predictive uncertainty. We decided to remove this part of introduction following the suggestion from referee #2 since our method and discussion did not further elaborate on this topic.

Comment 7; L. 57 – 58:

Strictly speaking, this is inaccurate. An ideal uncertainty quantification method should have ideal coverage of the quantiles. Imagine a 90% prediction interval, which consists of a 5% and 95% quantile. So, an ideal uncertainty quantification method should have a 5% and 95% coverage of the quantiles. If the 5% quantile is covered by 2% of the test samples, and the 95% quantile by 92% of the test samples, the PICP would be still 90%, even though the uncertainty was wrongly quantified. See for example the short paper on PICP from Pinson & Tatsu (2014) or figure below.



Literature

Pinson, P., & Tastu, J. (2014). Discussion of “Prediction intervals for short-term wind farm generation forecasts” and “Combined nonparametric prediction intervals for wind power generation”. IEEE Transactions on Sustainable Energy, 5(3), 1019-1020.

Reply: Thank you. We agree that the evaluation with PICP could be potentially biased, and evaluating the quantile is a way to address this issue. Methods suggested by Schmidinger and Heuvelink (2023), namely quantile coverage probability and probability integral transform, could address these issues better than PICP.

However, regarding the method used in this study, CP only provides a 90% interval instead of a predictive distribution. Predictive CDFs that are required for probability integral transform cannot be directly calculated. Since MCCP is just MC with extended upper and lower bounds, its prediction distribution is also derived from the original MC.

Indeed, quantile of CP can potentially be calculated, but it won’t be derived from a predictive distribution. We acknowledge the issue about one-sided bias and have a section in the 3.3 Limitations and future applications to discuss this issue.

“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, such as quantile coverage probability and probability integral transform, are thus needed to evaluate the uncertainty quantification in the future.”

Comment 8; L. 61 – 66

It is hard to understand why the authors mention only bootstrapping in this context. It is correct, that bootstrapping is used occasionally in soil science to quantify uncertainties. However, this is a common methodological error because bootstrapping only creates “confidence intervals” not “prediction intervals” i.e., pure bootstrapping was never intended to be used to quantify predictive uncertainties. On the other hand, the authors leave out the most commonly used method in soil: quantile regression (e.g., quantile regression forest, XGBoost with quantile loss function etc.) or even conformalized quantile regression. Quantile regression is not (yet?) well implemented for deep learning, which is why MC-CP becomes relevant but the introduction feels incomplete in the context of soil. More so, because conformalized quantile regression was recently introduced in soil by Kakhani et al. (2024), which is in its logic very related to MC-CP.

Literature

Kakhani, N., Alamdar, S., Kebonye, N. M., Amani, M., & Scholten, T. (2024). Uncertainty quantification of soil organic carbon estimation from remote sensing data with conformal prediction. Remote Sensing, 16(3), 438.

Reply: Thank you for the suggestions. We revised the 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 9; L. 76 – 86

The concept of aleatoric and epistemic uncertainty may be applied to this section.

Reply: We removed the introduction about aleatoric and epistemic uncertainty.

Comment 10; L. 93 – L. 95

One may argue that quantile regression could do so too, which is why it needs to be discussed somewhere earlier.

Reply: We added introduction about quantile regression in Comment 8.

Comment 11; L. 96

More appropriate would be to replace “we applied a strategy to increase the PICP of MC dropout” with “we applied a strategy to improve the PICP coverage of MC dropout”, unless MC generally leads to too narrow prediction intervals and not suboptimal coverage.

Reply: We revised it accordingly.

Comment 12; L. 118 – 120:

During my first read, I was wondering how the testing and validation was done. It is mentioned in L. 203 but for better readability it could be already defined here. I agree with it either way.

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. Only the training and validation data were used in building the model."

Comment 13; L. 123

It may be considered to use "it has been proven" instead of "it has been proved".

Reply: We corrected it accordingly.

Comment 14; L. 134:

Sounds slightly off. One suggestion to highlight that the predictive distribution is inferred from 100 trained CNN models with dropout layers: "In practice, a CNN model with dropout layers was trained 100 times to generate a predictive distribution".

Reply: Thank you for the comment. Monte Carlo dropout method generates predictive distribution by performing numerous (in this case 100) forward passes with the same CNN model with dropout activated.

From Gal and Ghahramani (2016): *"Note that the dropout NN model itself is not changed. To estimate the predictive mean and predictive uncertainty we simply collect the results of stochastic forward passes through the model."*

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."*

We modify this sentence to clarify this process:

"In practice, a CNN model with dropout layers was trained and performed 100 forward passes with dropout layers activated to generate a predictive distribution (Bethell et al., 2024)."

References:

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.

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.

Comment 15; L. 135-136:

The 90 % prediction interval is not the difference between the 5th and 95th quantile, but the prediction interval itself is the interval defined by those two quantiles. The difference is the “(mean) prediction interval width”. In Eq. 1 it is shown correctly, meaning it is just an issue of terminology.

Reply: Thank you for correcting this. We revise the content as follows:

“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 16; Eq. 1/ L.137:

C is the 90% prediction interval, which is relatively logical given the previous sentence. Nonetheless, it could be defined in the text. Also, if it is called C90, it would follow the scheme of $\hat{q}5$ and $q95$.

Reply: Thank you. We define that $C_{MC,\ 90}$ indicates the 90 % prediction interval in the text.

“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 17; L. 222:

Wording could be considered: it is correct that it is expectable that the model fails to do proper out-of-domain predictions, but the performance is not only "poor" but completely unusable since the R2 = -6.64. The word "poor" indicates to me a model with an R2 around or slightly above 0.

Reply: We agree. Combining the suggestion from referee #2, we modified the sentence as follows:

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

Comment 18; L. 250/Fig. 3:

It seems a bit incoherent that the plot covers the 0 – 90% range instead of 0 – 100% range.

Reply: Fig. 3 is now as follows:

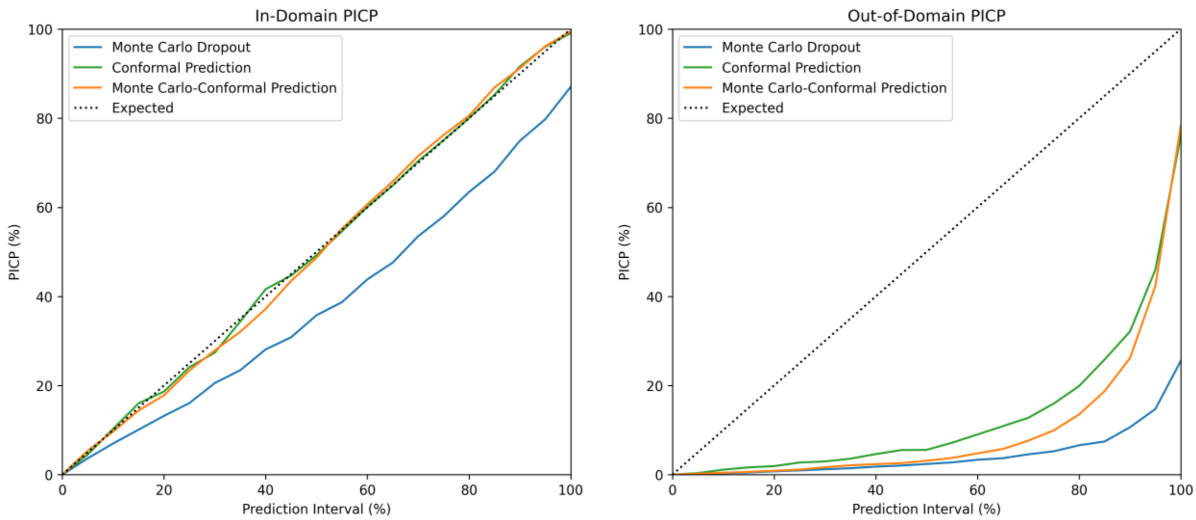


Figure 3: Prediction interval coverage probability (PICP) of in-domain and out-of-domain samples at different prediction intervals for Monte Carlo dropout, conformal prediction, and Monte Carlo-conformal prediction.

Comment 19; Out-of-domain results and discussion:

I do not fully agree with the discussion on the out-of-domain results for the following reasons: First of all, it is not clear why the authors do not show the PICP for the out-of-domain predictions, presumably because it has very poor coverage, given that the MPIW is only marginally larger even though the model is extremely bad for out-of-domain predictions. A much larger uncertainty (i.e., MPIW) should be expected here.

Instead, the authors focus on the fact that the MPIW is slightly larger for out-of-domain predictions with MC-CP and MC compared to in-domain predictions. It correctly shows that the model is somewhat aware that it is less certain for out-of-domain data. The authors see this as a reason to conclude that MC-CP is able to address the uncertainty of out-of-domain samples (L. 278-281). However, the results do not really support this claim because the MPIW alone is uninformative without information on the coverage. I highly assume that the MPIWs are still not wide enough. For extensive conclusions, the authors should include the coverage for the out-of-domain predictions. Hence, the coverage of the out-of-domain data needs to be included as well!

A second proceeding problem may occur if the PICP is used for evaluating out-of-domain samples, and it is strongly associated with the previous comment 7.

It can be expected that the observed values will be much more likely to be above the 95% quantile (as shown in the right example of Fig. 2!) and less often below the 5% quantile because the observed out-of-domain values have higher clay values than the model is trained on. Hence, evaluating the quantiles would make much more sense than using the PICP. This is the more standard practice and has been addressed in the context of soil too (Schmidinger & Heuvelink, 2023).

Literature

Schmidinger, J., & Heuvelink, G. B. (2023). Validation of uncertainty predictions in digital soil mapping. Geoderma, 437, 116585.

Reply: Thank you. We made necessary updates to address this comment.

1. We agree that the PICP of out-of-domain samples are also important for the context. We updated Fig. 3 and Table 4 to include the PICP of out-of-domain samples:

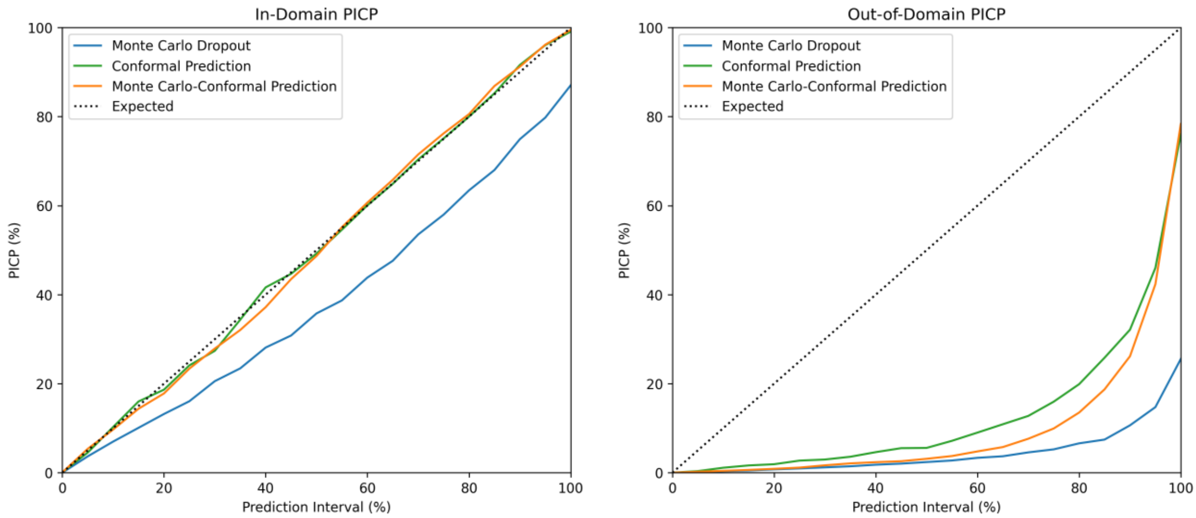


Figure 3: Prediction interval coverage probability (PICP) of in-domain and out-of-domain samples at different prediction intervals for Monte Carlo dropout, conformal prediction, and Monte Carlo-conformal prediction.

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

2. We acknowledge that the MPIW of out-of-domain samples was not dramatically different from the MPIW of in-domain samples. In our example, the MPIW of out-of-domain samples (10.43 %) was different from the MPIW of in-domain samples (9.05 %) by 1.38 %, which was about 15 % (1.38/9.05) increasement of width. 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 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, the difference of MIR spectra between mineral soils and organic soils could be bigger since this plot only includes soil samples with organic carbon less than 12%.

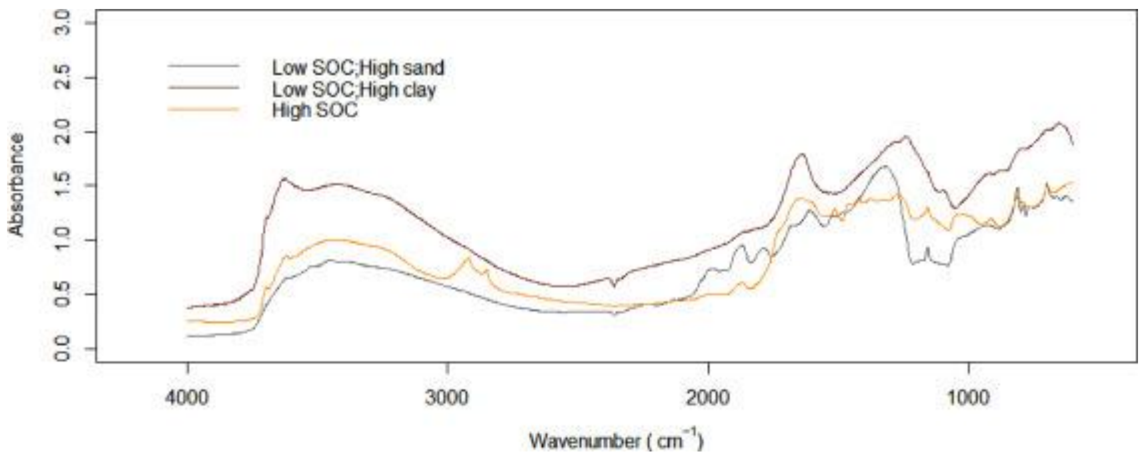


Figure from Ng et al. (2022)

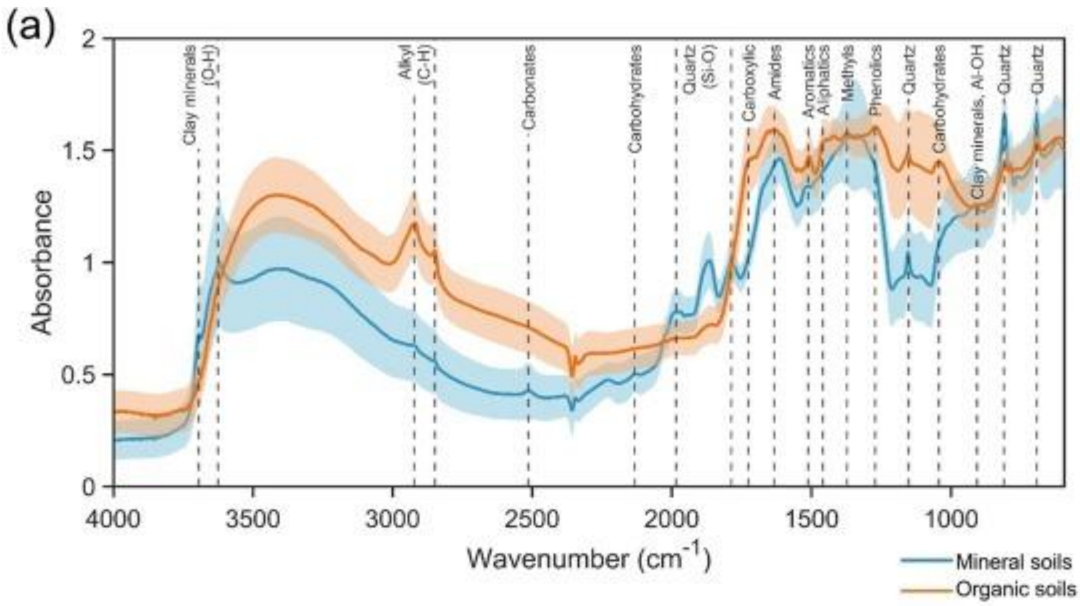


Figure from Zhang et al. (2022)

We included the above-mentioned discussion about the possible reasons that MC-CP did not generate a wider MPIW for out-of-domain samples:

"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."

3. Yes, we agree that the out-of-domain samples are more likely to be above the 95% quantile. This is reasonable since the out-of-domain samples have clay contents way higher than the in-domain samples, and the model always underpredict the clay content. We have a short discussion about the one-sided bias of out-of-domain samples:
   "In the present study, out-of-domain samples exhibited higher clay content than in-domain samples, with predicted values tending to be lower than the real value (Fig. 1). Consequently, observed values were more frequently above the 95 % quantile of prediction distribution, as illustrated by the out-of-domain example in Fig. 2. This one-sided bias arises from the separation of 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 20; L. 255:

MPIW instead of PIW.

Reply: Thank you. We corrected it accordingly.

Comment 21; L. 344:

"The authors benefited from the shared code of Daniel Bethell. The manuscript would become much more impactful for the soil community if the authors shared their code as well. This would increase the usability of the MC-CP method. Given that some KSSL data has been published in OSSL, it would also be easy to reproduce the study."

Reply: Thank you for the suggestion. We will make the code publicly available with the publication.