Dear Thorsten, thank you for the thorough revision of our manuscript. We address your comments below:

General remarks:

The paper is not easy to read (at least for a non-native speaker) and requires some restructuring and revision.

The text of the methodology section will be improved based on many comments that we received. We will clarify and explain terms that are not familiar to non-experts in deep learning.

The title starts with "Multi-source data integration". CNNs are explained in detail, but the integration of multi-source data is not adequately explained and discussed. The explanation of the CNNs lacks some detail.

We will try to expand and better describe the integration of multi data sources.

Why did you only use three covariables?

We used only three covariates because these were found to be the most important predictors in our case study at the fine spatial resolution. The same covariates are used in a recent paper of Somarathna et al (2018) https://doi.org/10.1016/j.scitotenv.2018.02.302

What is the difference between a window and a filter?

The difference between a window size and a filter will be better emphasized in the revised manuscript.

Specific comments:

Abstract:

The abstract is very condensed.
At the end you mention "different window size of input covariates matrix". This is not clear and needs an explaination.

We like the abstract but we will try to add a bit more detail on the results. We will rephrase the last sentences.

Introduction:

What was the purpose of DSM techniques before they were used to predict soil properties?

We will make small change in the text.

Can you explain this a little? The only point I really see is the size of the neighborhood.
Some approaches require less preprocessing than traditional terrain analysis. And you can simply set neighbourhood size as large as possible. For most approaches the resolution is the same for all input data. So I don't see any real "drawbacks" here. It's probably the same as setting up a CNN.

A bit more explanation will be provided.

In your paper you use three covariates. I assume that processing large spatial data sets with a CNN is a challenge. However, some of the other methods quoted here have no problem handling hundreds of covariables/scales. So not sure if they really have "drawbacks".

In view of your comment and those made by the other referee, we will and a short paragraph in the discussion to explain the limitation of our study using three covariates, and how CNN can handle more covariates.

What is the difference to approach presented by Padarian et al. (2018) especially when you write that "Padarian et al. (2018) have shown that it is possible to use CNN for soil mapping while accounting for contextual covariate information"?

The approach presented in our paper is an extension of the paper of Padarian (2018). Since using CNN is relatively new, we still need to develop approaches for mapping using CNN. In addition, Padarian (2018) use CNN for a country extent mapping, while we use a small, regional case study. We will mention it in the introduction.

Model definition:

Is this different with CNN?

Yes, in geostatistics, the measurements of the soil properties are a realization of an underlying Gaussian field. This is clearly not the case in deep learning, and generally in machine learning, where the measurements of the soil properties are spatially independent. We will add 'in space' to the sentence you mention.

Please explain the ReLU activation function and why you chose it.

ReLU activation function will be explained in the revised version of the manuscript.

CNN:

Small comments will be addressed directly in the revised manuscript. We agree with the reviewer comments and we will address them in the revised version. Most of them are also commented by the other referees.

There is no equation showing how this is done.

Regarding the case where we have c=3 covariates, we had this equation at an earlier version of the manuscript but we decided to not include it to not confuse the reader.

Figure 1: I suggest to split it in two separate figures. Does the convolution reduce the image size or just the pooling?

The authors would like to keep this figure as it is. We will make a better link with the text to explain each step.

Parameter estimation:

Is this "parameter estimation" or "learning (with backpropagation)"

Both are correct. We rather use the term 'parameter estimation', but learning could also be accepted.

Multi-source data integration:

The term "multi-source data integration" is in the title and is the interesting and important part of this article. However, this section does not show how it really works. How is the function updated? An example would be good.

We are not sure what the reviewer means by 'an example would be good'. We agree that we can add more explanation and we will do so in the revised version.

Quality of prediction:

I'm not sure if the formulas for RMSE are R2 are required.

We can indeed remove r2 and RMSE equations.

Why no cross-validation? Single subsets are generally not recommended for validation.

In our case study, it was not really possible to make a k-fold cross validation because of the computational load of calibrating a single CNN model. This is mentioned in the Discussion. Performing a cross-validation would provide more stable estimates of the validation statistics but would not change drastically the results. Splitting the dataset between validation and calibration sets is very common and widely accepted in soil mapping. The aim is just for a comparison with another method (RF), not for uncertainty estimates. This splitting will not benefit or disadvantage any method as the same dataset is used.

Why only 1000 trees? I suggest setting it to at least 2000 trees and then optimizing mtry and nodesize. However, this is not crucial as it will not significantly increase the prediction accuracy of RF in this study. The difference is the lack of spatial context.

The parameters of the RF model have already been tuned manually. Based on another reviewer comment, we will tune the parameter using a search algorithm for the revised version of the manuscript.

Why 90:10 in this case?

90:10 is an arbitrary choice, but this is common in the deep learning literature.

What "window size"? Filter size? In Table 1 I find 3x3.

The difference between window size and filter size will be better explained in the revised manuscript.

Why is the "window size" optimized separately? What were the inital settings of the hyperparameters when optimizing the window size?

The window size is optimized separately because it relates to the number of contextual information we provide to the model. This is not a model parameter not a hyperparameter of the model architecture.

Figure 6:

The maps look noisy and probably show artificial horizontal and vertical stripes that have nothing to do with the input data and look like artifacts from the convolution. But without examining the data, this is only a guess.

The map is indeed a bit noisy but we cannot see the vertical and horizontal stripes.

Discussion:

Can you produce uncertainty maps reflecting influence of the measurement error?

Producing uncertainty maps would certainly provide a valuable extension of our work. However, this is not straightforward. Padarian (2018) provides a maps of confidence intervals based on

predictions from several CNN models calibrated using a bootstrap sample of the training data. This refers to uncertainty due to the model, and does not include data variance. To obtain prediction intervals (and so the total variance of the prediction), we also need to account for the data noise variance (see second term of Eq. 3 in the paper of Khosravi et al., 2011). This is not simple and deserves an important research effort. In addition, this is not sure that measurement error can also be included because deriving prediction intervals requires another objective function instead of the MSE.

What is the effect on the final maps as well as prediction accuracy when setting all weights to 1?

We believe there is no reason why we should set the same weights as we already know that the measurements have different uncertainty. It is not a meaningful exercise, because we already know that the quality of data is different.

I suggest to put more emphasis on this.

We will put more emphasize on the novelty of our work.

Can you explain the relationship between the window size, the size of the input image, and the amount of contextual information? A figure would be good.

This is closely related to the Fig. 1. Based on your comments and other referee comments, we will put more link between the text and this Figure.

This is not what is shown in the ConMap paper quoted by you, but the multi-scale one in which we have tested comparable window sizes. Multi-scale digital terrain analysis and feature election for digital soil mapping Geoderma 155 (3-4), 175-185).

Thank you for spotting this mistake. We will add the correct reference in the revised version of the manuscript.

Reference:

Khosravi, A., Nahavandi, S., Creighton, D., & Atiya, A. F. (2011). Comprehensive review of neural network-based prediction intervals and new advances. IEEE Transactions on neural networks, 22 , 1341–1356.