We thank Dr. Amelie Beucher for the constructive comments on our paper. You can find below our answer to your suggestions.

Concerning the ANN section:

Each layer contains units: nodes for the input layer (as no computation occurs at their level) and neurons for the hidden and output layers;

We agree with this comment.

You should avoid confusion on the term "model parameters" right in this ANN section. Further in the case study part (3.2.2 Parameter estimation), you make the difference between model parameters and "the hyperparameters of the model architecture": these should be clearly defined before. I would first consider that the model parameters are the numbers of hidden layers, hidden neurons and iterations (i.e. parameters of the model architecture), not the connection weights and bias. You need to clarify this.

We call the connection weights and the bias as 'model parameters' as opposed to with the 'hyperparameters of the model architecture', i.e. number of layer, number of neurons per layer, etc... The authors would like to keep the names as it is because the model parameters have to be estimated by minimization of the objective function, while the hyperparameters of the model architecture can simply be 'tuned'. We will emphasize the differences between the two earlier in the manuscript to avoid confusion.

You use the abbreviation ReLU (for Rectified Linear Unit): you should clarify it and shortly explain it (giving the equation is not enough).

We agree, we will explain the ReLU function in the revised manuscript.

Equations 3 to 5 are not self-sufficient; you need to introduce them more. Particularly, you start by writing "For k = 1, : : :, L hidden layers,", but h0 is for the input layer and hL for the output layer. You have to modify Figure 1 accordingly as well.

Thank you for spotting this small mistake, we will modify Fig. 1 accordingly.

Concerning the CNN section:

You should refer more to your figure 1 and list shortly the different steps or layers of a CNN before describing them one by one.

We will modify the text to refer more to Fig. 1. This comment has also been suggested by other reviewers.

Define the word convolution straight ahead (i.e. element-wise product and sum between two matrices), this is a new concept for most in the DSM community.

We will include this sentence in the revised manuscript.

You use the Max-pooling operator and should define it simply (i.e. selecting the maximum value in the convoluted image using a window size). For now, we only read the term in table 1.

In Table 1, we mentioned max-pooling and described its meaning and operation in the text . In the revised manuscript we will add a sentence to explain the max-pooling operation.

You should clarify what you mean by "the number of channels (aka depth)": Which depth are you referring to?

We will modify the text accordingly to your comments.

You do not define the flatten operation as such, only writing that ": : :the last convolution returns an image of size 1 x 1 and with a number of channels. This is a vector that we can pass to a fully connected layer." Again, we only read the term in table 1.

The flatten operation will be better described in the revised manuscript.

You should state clearly that a fully-connected layer is an ANN. For now, it is only noted in figure 10s caption.

We define the ANN in the section 2.: 'We first describe the principle of Neural Networks (NN), basis of CNN'. We will refer more to Fig. 1 to make it clear.

You neither define the dropout operation. It should be in the methodology.

We will explain it in the methodology section.

Concerning the parameter estimation:

You define the dataset D as ": : : a 4-D matrix of size n x c w x h : : :", but you only defined n (i.e. the number of calibration points) much later in the paper.

The number of sampling location n is defined at the beginning of the paper, line 15.

You have to be clear and consistent when you use the term "weights" (in this section and the whole paper): the measurement error weights (defined in the 2.5 section) are different from the connection weights (as in the model parameters theta). Using only the term "weights" is confusing.

We will make clear the distinction between the two in the revised manuscript.

It is important that you describe shortly the Bayesian optimization here and not in the case study part (3.2.2 Parameter estimation);

While we may agree on this comment, we would like to keep the Bayesian optimization paragraph in the practical implementation section. This is simply a tuning procedure that we decided to perform but we might confuse the reader by adding it to the methodology section. Therefore, we prefer to keep the paragraph on Bayesian optimization in the implementation section.

You should also clarify the backpropagation concept in connection with the error gradient descent. These are crucial points for neural networks in general.
We will add a few sentences about it.

Comments on the manuscript:

Most comments are straightforward to address. Some small answers below:

We will add a north arrow to the figures as well as a scale.

Concerning the addition of a plot illustrating the correlation between auto-correlation range and window size, we rather do not include it in the article. This paragraph in the discussion is a speculation that requires further investigation (as it is already mentioned).