We thank the editor for the additional comments. Our response is color-coded in blue.

I understand that this study is focused on proof of concept for the two example regions. However, given the random selection of the test dataset, can you be sure that the model really does generalise well to other meteorological conditions in these two regions (not other regions)? For example, the simultaneous footprints in the files

X_Y_-97.94Ex33.03N_2013102000.npz

X_Y_-97.98Ex33.05N_2013102000.npz

are for receptors that are close together, so the corresponding footprints and inputs are very similar. My understanding is that one of them may have been in the training dataset and the other in the test dataset (if this doesn't apply to these two footprints, there are certainly other examples). In this case, the model only needs to reproduce the footprint in the training dataset during testing, which would lead to good test results even for a non-generalising model. Could you discuss whether or not this might be a problem here? It could be tested and avoided by, for example, splitting the dataset into different time periods rather than randomly.

We appreciate the editor for this comment. Yes, we acknowledge that similarities between samples are hard to fully avoid during the construction of version 1 of FootNet, even if the construction of the training and test data sets is done by splitting data by time periods rather than random selection. We tested the training split proposed by the editor using the training data set, and indeed found degradation in the performance.

We also evaluate the training split strategy proposed by the editor in our more recent updates of the model. We find little-to-no difference in performance, indicating that when the data volume is largely increased the generalizability of FootNet could be improved.

We have added the following line to the text to emphasize the results demonstrated from the construction of FootNet version 1 could be affected by different split strategies of the training data set:

Line 138: “However, it is worth mentioning here that we find some performance degradation using an alternative splitting of the data based on different time periods. Because the training data set used to construct version 1 of FootNet has a relatively small size, similarities between samples are hard to fully avoid by randomly selecting training data samples, which could lead to generalizability issues when using FootNet version 1 over regions and time periods too different from the training data set. This generalizability issue could be largely mitigated by increasing the volume of the training data set in the future (Dadheech et al., 2024).”

One minor comment: in footnet.py, the footprint transformed by the protocol seems to be shifted by 20, which is not mentioned in the manuscript and corresponds to a scaling of the footprint by exp(20).

Thank you. The offset is applied to filter and remove too small footprint values generated by the physics-based model. We have now mentioned this in the manuscript.

Line 80: “The transformed footprints are filtered to remove values smaller than -20 and then shifted by +20, corresponding to a scaling of the raw footprints by exp(20).”