EXPLORING THE POTENTIAL OF NEURAL NETWORKS FOR SPECIES DISTRIBUTION MODELING

Robin Zbinden*, **Nina van Tiel***
EPFL, Switerzland
firstname.lastname@epfl.ch

Benjamin Kellenberger
Yale University, USA
bak45@yale.edu

Lloyd Hughes, Devis Tuia
EPFL, Switerzland
firstname.lastname@epfl.ch

ABSTRACT

Species distribution models (SDMs) relate species occurrence data with environmental variables and are used to understand and predict species distributions across landscapes. While some machine learning models have been adopted by the SDM community, recent advances in neural networks may have untapped potential in this field. In this work, we compare the performance of multi-layer perceptron (MLP) neural networks to well-established SDM methods on a benchmark dataset spanning 225 species in six geographical regions. We also compare the performance of MLPs trained separately for each species to an equivalent model trained on a set of species and performing multi-label classification. Our results show that MLP models achieve comparable results to state-of-the-art SDM methods, such as MaxEnt. We also find that multi-species MLPs perform slightly better than single-species MLPs. This study indicates that neural networks, along with all their convenient and valuable characteristics, are worth considering for SDMs.

1 INTRODUCTION

Describing and understanding the geographic distribution and environmental suitability of species is a central question in ecology and biogeography. It has become increasingly meaningful in the face of anthropogenic pressure on biodiversity (Barnosky et al., 2011) and has occasioned the development of species distribution models (SDMs)[1]. Such models relate species occurrence data with environmental variables and are used to understand and predict species' distributions across landscapes (Elith and Leathwick, 2009). Such predictions may allow the identification of areas likely to be invaded by non-native species (Thuiller et al., 2005b), the forecast of shifts in distribution due to climate change (Thuiller et al., 2005a), or the selection of field survey areas to speed up the discovery of a new population of a species (Fois et al., 2015). SDMs can therefore be used to support decision-making for species conservation (Guisan et al., 2013) and prevent biodiversity loss, which is generally beneficial for the climate as shown by Shin et al. (2022).

However, species occurrence data used for SDMs are expensive and complicated to collect. Datasets are typically small, especially for rare species and in regions that are less well-researched, restricting their predictive power and making them difficult to transfer to other regions. These datasets consist of either *presence-absence* data from systematic field surveys, or *presence-only* data, which is easier to collect and more widely available, especially with crowdsourcing initiatives such as iNaturalist (Van Horn et al., 2018) or Pl@ntNet (Joly et al., 2016). When using presence-only data, *pseudo-absences* or *background points* may be generated (randomly or according to specific sampling strategies) and used as negative samples in the model (Phillips et al., 2009).

*Indicates equal contribution

[1]It should be noted that many names exist for this type of model. In some contexts, *ecological niche model* may be more appropriate to refer to these types of models (Peterson and Soberón, 2012). For simplicity, we will use the term *species distribution model*.

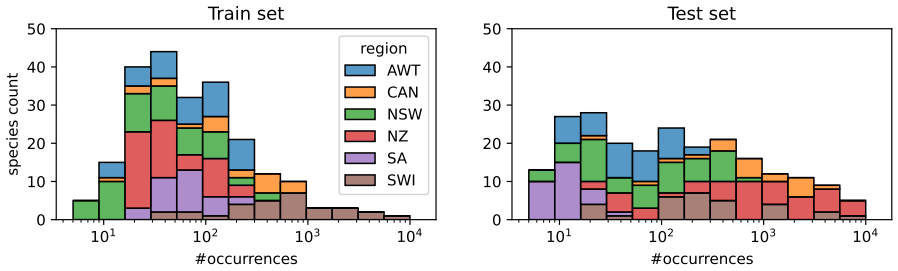

Figure 1: Number of occurrences per species in the training and test set for each region. The x-axis is logarithmic. The two distributions are quite different, but both follow a long-tailed distribution.

Well-established methods for SDMs are adapted to the customarily small datasets and include simple statistical models, such as generalized linear models (GLMs) and generalized additive models (GAMs), as well as machine learning methods like random forests, support vector machines, and MaxEnt (Phillips et al., 2017; Zhang and Li, 2017; Valavi et al., 2022). Generally, species are modeled independently of each other, despite the fact that biotic interactions shape species' distributions at local scales (Wisz et al., 2013). Deep learning methods are revolutionizing ecological modeling (Tuia et al., 2022) and might be particularly suitable in an SDM context: they can model highly non-linear relationships between environmental variables and distributions, and can elegantly model multiple species at once. Hence, with the growing amount of available data, deep learning algorithms for SDMs have begun to be explored (Botella et al., 2018; Deneu et al., 2021; Lorieul et al., 2022), but are far from being standardized in the field of ecology.

In this work, we explore the performance of neural networks, specifically multilayer perceptrons (MLPs), for SDM using tabular data from a benchmark dataset (Elith et al., 2020). We compare our results to the benchmark evaluation of well-established methods (Valavi et al., 2022). Furthermore, we compare a single-species approach, i.e., a set of MLPs trained individually for each species, to a multi-species approach, where a single MLP is trained on the data for a set of species and therefore performs multi-label classification. With both approaches, we achieve comparable results to state-of-the-art methods such as MaxEnt. The multi-species model performs overall slightly better than the single-species model, particularly for species with very few records. We hypothesize these species may benefit from the integration of other related species, as the model could recognize common environmental patterns. These results pave the way to ecology-guided neural network strategies for SDMs exceeding the simple ones presented in this short paper.

2 Model and Methods

Dataset: We use the dataset from Elith et al. (2020) which contains occurrence data for 226 species[2] from 6 regions of the world: Australian Wet Tropic (AWT), Canada (CAN), New South Wales in Australia (NSW), New Zealand (NZ), South America (SA), and Switzerland (SWI). For each region, we have access to 11 to 13 different environmental covariates, including climatic and pedological variables. This dataset was made public, albeit with anonymized species, as a benchmark, so that different methods can be tested and compared (Elith et al., 2006; Phillips et al., 2017; Valavi et al., 2022). It is particularly valuable because it contains an independent test set made of presence-absence data, which allows unbiased evaluation of the SDM results. In contrast, the training set consists of presence-only data and background points, uniformly sampled for each region, as is the norm for observational data for SDMs. To match our setup to that of Valavi et al. (2022) and be able to compare our results, we also use the $50,000$ background points provided by the authors, rather than those provided in the original dataset. The training data is representative of what is usually available for SDMs and is characterized by a long-tailed distribution over the number of occurrences per species. The distribution of samples per species can be observed in Figure 1: 50% of the species have less than 60 occurrences in the training set. Furthermore, we can also note a mismatch between the distributions of the two sets, indicating that the data was collected in different ways.

[2]Like in Valavi et al. (2022), we exclude one species for which only 2 occurrences were available in the training set. Therefore, we consider 225 species.

	AWT	CAN	NSW	NZ	SA	SWI
MaxEnt	**0.686**	0.584	0.713	0.738	0.804	0.809
XGBoost	0.653	0.568	0.706	0.720	0.788	0.815
Random Forest	0.675	0.572	0.718	0.746	**0.813**	**0.818**
Ensemble	0.683	0.580	**0.723**	**0.749**	0.806	0.812
Single-species MLP	0.666	0.589	0.688	0.715	0.799	0.808
Multi-species MLP	0.617	**0.605**	0.708	0.714	0.803	0.815

Table 1: Mean AUROC across species for each region achieved by our models and the state-of-the-art models[3] of Valavi et al. (2022). The results of our single- and multi-species models are averaged over 10 different random models initialization to add statistical power.

Model architecture: We use a multi-layer perceptron (MLP), a well-studied neural network architecture, which can be repurposed from a single-species to a multi-species model by simply increasing the number of neurons in the output layer. We take inspiration from the MLP mixer architecture of Tolstikhin et al. (2021), which incorporates recent advances in deep learning. Accordingly, one layer of our MLP consists of the sequential application of batch normalization (Ioffe and Szegedy, 2015), a fully connected layer, the SiLU activation function (Elfwing et al., 2017), and dropout Srivastava et al. (2014). It is repeated L times with skip connections, where L is a hyperparameter. A sigmoid function is applied to each neuron in the last layer to obtain a prediction between 0 and 1 that can be interpreted as a species occurrence score. An illustration of the architecture of our MLP models can be found in Appendix A.

Training: MLPs are known to be highly sensitive to hyperparameters, making it difficult to regularize them properly (Grinsztajn et al., 2022). We use the Optuna library (Akiba et al., 2019) to improve and automate the hyperparameter search for both the multi-species (one global model) and the single-species (one model per species) cases. To account for the class imbalance, we use a weighted binary cross-entropy, where the weight for each species is the ratio of the number of presences to the number of background points. In the case of the multi-species model, we take into account the number of background points plus the number of sites where that species was not observed. Further implementation details can be found in Appendix A.

3 Results

We evaluate our models via the Area Under the Receiver Operating Characteristic curve (AUROC) for each species and the average AUROC across species for each region. In Table 1, we see that, although no model systematically outperforms all others, both our single- and multi-species MLPs achieve AUROC values that are comparable to those of the state-of-the-art SDM methods evaluated in Valavi et al. (2022). The left panel of Figure 2 shows that the AUROC values vary a lot across and within each geographical region for our methods, as well as for MaxEnt.

Our multi-species MLP outperforms the single-species models in four regions. We observe in the right panel of Figure 2 that species with a low number of occurrences in the training set seem to be more strongly affected by the choice of the model. Species with sufficient occurrence data (>200) show very similar performance between single- and multi-species models, whereas species with very few occurrences (<10) perform on average better with the multi-species model, showing the inherent class regularization effect of a multi-species model, which indirectly leverages co-occurrences information.

4 Discussion

Table 1 shows that no model outperforms any other consistently. Although our MLP models perform comparably, state-of-the-art SDM methods, notably tree-based methods like random forest, remain among the top models for SDM. In fact, tree-based methods generally handle tabular data better

[3]We consider the three methods obtaining the best mean AUROC over all species in Valavi et al. (2022), i.e., MaxEnt, Random Forest down-sampled, and Ensemble. We also add XGBoost as it is often considered to be a state-of-the-art model for tabular data.

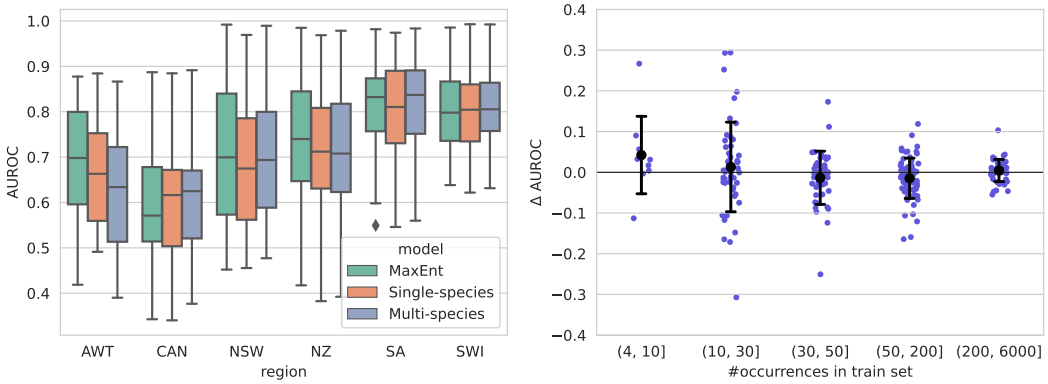

Figure 2: **Left**: Box plots showing the variability of AUROC across species within each region, for MaxEnt (Valavi et al., 2022) and our models. **Right**: Difference of AUROC between our multi- and single-species models for each species, grouped by the number of occurrences in the train set. Points above the horizontal line indicate a species where the multi-species model obtains a higher AUROC than the single-species model and vice-versa. Error bars represent the standard deviation along with the mean over the species of the given group.

than neural-based approaches (Borisov et al., 2021; Grinsztajn et al., 2022). This is consistent with our results, particularly when comparing single-species MLPs with random forests. SDMs almost always use tabular data, despite the fact that species distributions are inherently geospatial. One way to account for the geospatial nature of the task is to consider raster images of environmental variables or satellite images as input features (Deneu et al., 2021; Kellenberger et al., 2022). Such data naturally informs about spatial context and patterns beyond the variables observed at the specific location of the occurrence, leading to a natural spatial regularization. Convolutional neural networks are more adapted to image or raster input data than tree-based approaches or MaxEnt.

Furthermore, neural networks are more flexible for integration into more complex machine learning systems and are a promising approach to improve upon the current state-of-the-art SDM models, for instance by incorporating other modalities alongside tabular data, or including expert knowledge, such as information about species interactions, through knowledge-guided machine learning (Karpatne et al., 2022). Even when remaining in the realm of tabular data, our results show that including other species in the training can improve the model performance. As shown in Figure 2, species with very few occurrences seem to particularly benefit from the co-occurring environmental patterns of other species seen during training. Moreover, considering one multi-species model instead of numerous single-species models reduces the computational burden and complexity of handling as many models and hyperparameter searches as there are species. A multi-species framework would also facilitate the integration of biotic interactions. Currently, species interactions are rarely considered in SDMs, despite the fact that they may have an important influence on species distributions (Guisan and Thuiller, 2005; Wisz et al., 2013). Pre-training approaches, such as fine-tuning a multi-species initialization for a single-species model, or transfer learning from well-sampled species to related species with fewer recorded occurrences, remain possibilities to be explored. Note that the anonymization of the species in the Elith et al. (2020) dataset prevents us from identifying such related species and hinders us from investigating these avenues with this dataset.

5 Conclusion

In this work, we compare MLP models to well-established species distribution modeling methods and obtain results that are comparable to state-of-the-art methods on a standard benchmark where evaluation is properly handled, i.e., including real species absence data. We highlight the convenience and ecological significance of building SDMs that consider multiple species simultaneously, rather than a single species. Finally, we underline the potential of neural networks in SDMs, due to their ability to be integrated into more complex machine learning systems, such as knowledge-guided machine learning and pre-training approaches.

REFERENCES

T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

A. D. Barnosky, N. Matzke, S. Tomiya, G. O. Wogan, B. Swartz, T. B. Quental, C. Marshall, J. L. McGuire, E. L. Lindsey, K. C. Maguire, et al. Has the earth's sixth mass extinction already arrived? *Nature*, 471(7336):51–57, 2011.

V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci. Deep neural networks and tabular data: A survey. *arXiv preprint arXiv:2110.01889*, 2021.

C. Botella, A. Joly, P. Bonnet, P. Monestiez, and F. Munoz. A deep learning approach to species distribution modelling. *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*, pages 169–199, 2018.

B. Deneu, M. Servajean, P. Bonnet, C. Botella, F. Munoz, and A. Joly. Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLOS Computational Biology*, 17, 04 2021. doi: 10.1371/journal.pcbi.1008856.

S. Elfwing, E. Uchibe, and K. Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning, 2017. URL `https://arxiv.org/abs/1702.03118`.

J. Elith and J. R. Leathwick. Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution and Systematics*, 40(1):677–697, 2009.

J. Elith, C. H. Graham, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, et al. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2):129–151, 2006.

J. Elith, C. Graham, R. Valavi, M. Abegg, C. Bruce, S. Ferrier, A. Ford, A. Guisan, R. J. Hijmans, F. Huettmann, et al. Presence-only and presence-absence data for comparing species distribution modeling methods. *Biodiversity informatics*, 15(2):69–80, 2020.

M. Fois, G. Fenu, A. C. Lombrana, D. Cogoni, and G. Bacchetta. A practical method to speed up the discovery of unknown populations using species distribution models. *Journal for Nature Conservation*, 24:42–48, 2015.

L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

A. Guisan and W. Thuiller. Predicting species distribution: offering more than simple habitat models. *Ecology letters*, 8(9):993–1009, 2005.

A. Guisan, R. Tingley, J. B. Baumgartner, I. Naujokaitis-Lewis, P. R. Sutcliffe, A. I. Tulloch, T. J. Regan, L. Brotons, E. McDonald-Madden, C. Mantyka-Pringle, et al. Predicting species distributions for conservation decisions. *Ecology letters*, 16(12):1424–1435, 2013.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR.

A. Joly, H. Goëau, J. Champ, S. Dufour-Kowalski, H. Müller, and P. Bonnet. Crowdsourcing biodiversity monitoring: how sharing your photo stream can sustain our planet. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 958–967, 2016.

A. Karpatne, R. Kannan, and V. Kumar. *Knowledge Guided Machine Learning: Accelerating Discovery Using Scientific Knowledge and Data*. CRC Press, 2022.

B. Kellenberger, E. Cole, D. Marcos, and D. Tuia. Training techniques for presence-only habitat suitability mapping with deep learning. In *IEEE International Geoscience and Remote Sensing Symposium, IGARSS*, Kuala Lumpur, Malaysia, 2022.

T. Lorieul, E. Cole, B. Deneu, M. Servajean, P. Bonnet, and A. Joly. Overview of geolifeclef 2022: Predicting species presence from multi-modal remote sensing, bioclimatic and pedologic data. In *CLEF 2022-Conference and Labs of the Evaluation Forum*, volume 3180, pages 1940–1956, 2022.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

A. Peterson and J. Soberón. Species distribution modeling and ecological niche modeling: Getting the concepts right. *Natureza e Conservação*, 10:1–6, 12 2012. doi: 10.4322/natcon.2012.019.

S. J. Phillips, M. Dudík, J. Elith, C. H. Graham, A. Lehmann, J. Leathwick, and S. Ferrier. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological applications*, 19(1):181–197, 2009.

S. J. Phillips, R. P. Anderson, M. Dudík, R. E. Schapire, and M. E. Blair. Opening the black box: An open-source release of maxent. *Ecography*, 40(7):887–893, 2017.

Y.-J. Shin, G. F. Midgley, E. R. Archer, A. Arneth, D. K. Barnes, L. Chan, S. Hashimoto, O. Hoegh-Guldberg, G. Insarov, P. Leadley, et al. Actions to halt biodiversity loss generally benefit the climate. *Global change biology*, 28(9):2846–2874, 2022.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56): 1929–1958, 2014.

W. Thuiller, S. Lavorel, M. B. Araújo, M. T. Sykes, and I. C. Prentice. Climate change threats to plant diversity in europe. *Proceedings of the National Academy of Sciences*, 102(23):8245–8250, 2005a.

W. Thuiller, D. M. Richardson, P. Pyšek, G. F. Midgley, G. O. Hughes, and M. Rouget. Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. *Global change biology*, 11(12):2234–2250, 2005b.

I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021.

D. Tuia, B. Kellenberger, S. Beery, B. R. Costelloe, S. Zuffi, B. Risse, A. Mathis, M. W. Mathis, F. van Langevelde, T. Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1):792, 2022.

R. Valavi, G. Guillera-Arroita, J. J. Lahoz-Monfort, and J. Elith. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecological Monographs*, 92(1):e01486, 2022.

G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

M. S. Wisz, J. Pottier, W. D. Kissling, L. Pellissier, J. Lenoir, C. F. Damgaard, C. F. Dormann, M. C. Forchhammer, J.-A. Grytnes, A. Guisan, et al. The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological reviews*, 88(1):15–30, 2013.

J. Zhang and S. Li. A review of machine learning based species' distribution modelling. In *2017 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII)*, pages 199–206. IEEE, 2017.

A IMPLEMENTATION DETAILS

	Range	AWT	CAN	NSW	NZ	SA	SWI
#Layers	{2,...,6}	4	5	4	5	6	5
MLP width	{256,...,2048}	1327	285	456	465	1140	420
Learning rate	[1e-5, 1e-2]	4e-3	1e-4	2e-5	2e-3	2e-5	4e-4
Weight decay	[1e-5, 1e-2]	3e-3	8e-5	2e-3	9e-4	8e-5	9e-3
Dropout	[0.0, 0.3]	0.17	0.018	0.15	0.28	0.04	0.018

Table 2: Range and best hyperparameters found by Optuna per region for the multi-species model. The learning rate and the weight decay are sampled from the range in the log domain. The hyperparameters range is the same for the single-species model.

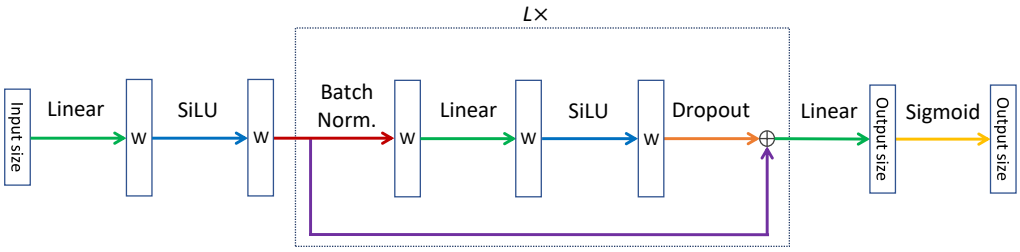

Figure 3: The architecture of our MLP models. The input size is the number of environmental covariates for a given region (a little more if the region contains one-hot encoded categorical variables). The size of the output is one for the single-species model, while it is the number of species in the region for the multi-species model.

We ran Optuna for 50 iterations for each model to find the best set of hyperparameters within the ranges shown in Table 2. For the multi-species model, each iteration of Optuna trains a model for 200 epochs, and we select the one that gives the best AUROC on the validation set (20% of the training set sampled uniformly at random) with early stopping. The best-performing hyperparameter-set for each region can be found in Table 2. For the single-species model, we use k-fold cross-validation, where k equals 3, as the number of presences for some species is very low. For the same reason, we use the entire training set for the training of the final single-species model, which prevents us from using early stopping. We therefore only train it for 50 epochs to avoid overfitting. Each iteration of Optuna considers the mean AUROC over the k folds to determine the best set of hyperparameters. We decided not to list the best set of hyperparameters for each of the 225 species.

We train all our models using the AdamW optimizer (Loshchilov and Hutter, 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon =$ 1e-8, a learning rate scheduler with an exponential decay of 0.95, and a batch size of 256. Some regions contain a few categorical variables, which take a small number of different values. We therefore one-hot encode them before feeding them into our MLPs. Finally, the architecture of our MLP models is illustrated in Figure 3. After finding a good set of hyperparameters, each model was trained with 10 different random initialization to add statistical power.