Species distribution modeling with time series data and deep learning

***Austin M. Smith*[1], *Cesar Capinha*[2], *Andrew M. Kramer*[1]**

[1]*Department of Integrative Biology, University of South Florida*, 4202 E. Fowler Avenue, Tampa, FL 33620, USA

[2]Centre of Geographical Studies, Institute of Geography and Spatial Planning, University of Lisbon, Rua Branca Edmée Marques, 1600-276 Lisboa, Portugal

INTRO:

Species distribution models (SDMs) are widely used to gain ecological understanding and guide conservation decisions. These models are developed with a wide variety of procedures - from regression-based approaches to mores to machine learning algorithms. A wide range of techniques have been studied to optimize model performance (Norberg et al., 2019; Velavi et al., 2022), but one property they nearly all share is the use of summary predictors (e.g., mean, standard deviation, etc.) that strongly simplify the temporal variability of driving factors. On the other hand, recent architectures of deep learning known as convolutional neural networks (CNNs) allow dealing with fully explicit spatiotemporal dynamics, thus removing the need to simplify the temporal and spatial dimension of environmental predictor data (Capinha et al, 2021; Smith et al. 2022; Ceia-Hasse et al. 2023). Given the apparent conceptual advantages of time-series-based CNNs over conventional approaches in the development of SDMs, a robust comparison of the predictive performance of the two approaches is needed.

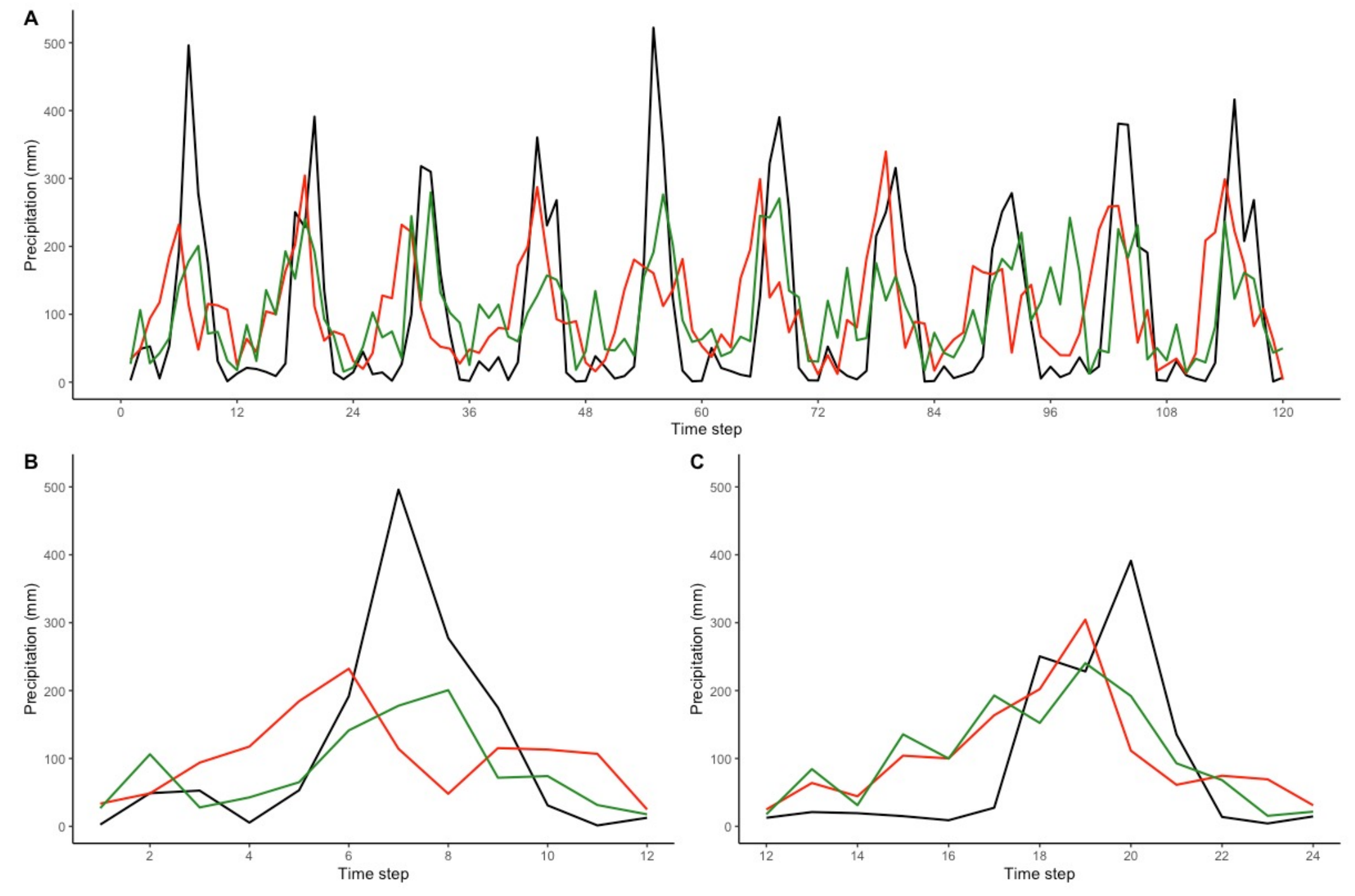

Figure 1: Example of time series dynamics for monthly precipitation . A.) Precipitation data from the 1990 - 2000 for three locations that fall along the same lines of latitude. The locations include Tampa, FL, US (green line), Nainapur, Uttar Pradesh, India (black line), and Songtao Miao Autonomous County, Tongren, Guizhou, China (red line). These points were extracted from WorldClim BIO12, which measured averaged annual precipitation. All three points have roughly the same pixel value of 1260 mm. B.) precipitation for the year 1990 and C.) 1991. Note, precipitation not only differs spatial, but also temporally.

Model	Abbreviation	Overview
Gradient boosting machine	GBM	An additive ensemble method that uses gradient boosting. This process involves generating a series of weak predictive decision trees built from a subset of available input variables, then calibrating hyperparameters on new trees to minimize the loss function.
Maximum entropy	MaxEnt	Creates a probability distribution function, using occurrence records, with the greatest entropy or spread (closest to uniform). The use of environmental background points creates an environmental gradient space. Occurrence record probabilities are then compared to the environmental space to rank suitability.
Random forest	RF	An ensemble method that generates a series of fully grown, unpruned decision trees, constructed from bagging (i.e., bootstrap aggregation). For regression models, the model output is the average across all trees and the majority class is the output for classification models.

Table 1: Set of popular conventional machine learning algorithms used as baseline examples .

METHODS:

Species records were collected from the National Center for Ecological Analysis and Synthesis (NCEAS; Elith et al., 2020) and include 85 anonymized species from Ontario, Canada, New Zealand, South America, and Switzerland. Climate and elevation data were extracted from Worldcim (Fick & Hijmans, 2017). CNN models were generated using the automated machine learning (AutoML; He et al., 2021) assemblage Python package McFly (Van Kuppevelt et al., 2020). 20 candidate models were generated and trained for five epochs. The best candidate model was chosen using area under the receiver operating receiver characteristic curve (ROC), then trained on larger collection of samples for up to 100 epochs or until monitored metrics stop improving (i.e., early stopping). Finally, fully-tuned models were tested on an independent presence-absence data set. Note, because of class imbalance in the data and an emphasis on predicting true occurrences in SDM procedures, area under the precision-recall curve (PR) was also measured (Sofaer et al., 2019). These models were then compared to the baseline models (Table 1)

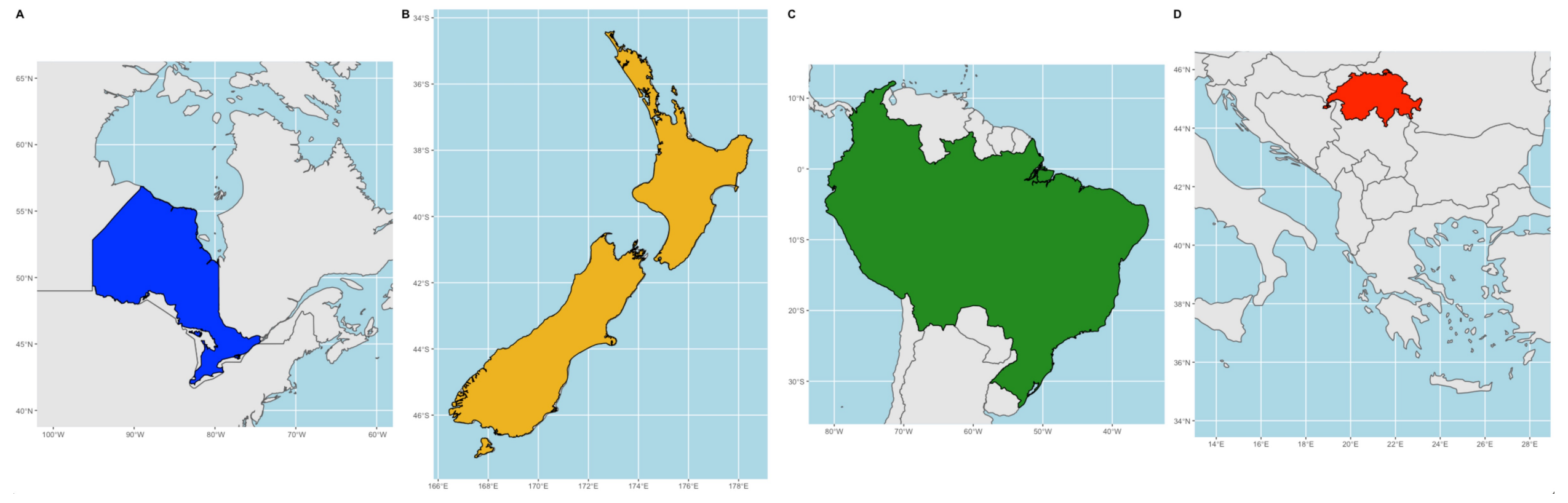

Figure 2: Regions where species occurrences were collected: A.) Ontario, Canada; B.) New Zealand; C.) South America: and D.) Switzerland.

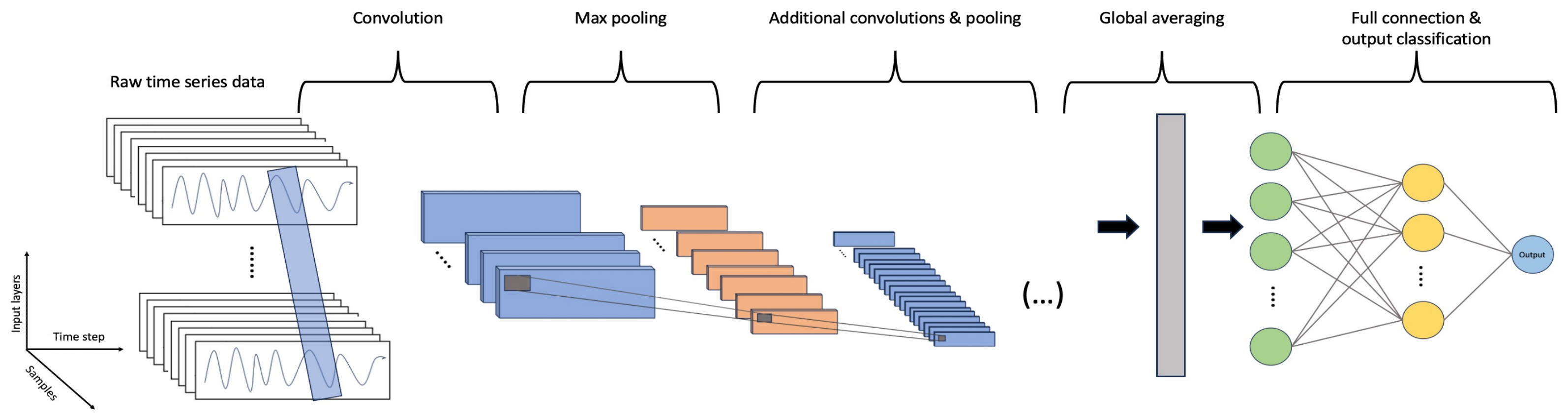

Figure 2: Convolutional neural network for time series classification**.** As the name implies, CNNs refer to convolutional layers; internal filter functions of varying length convolve with patches of the time series data to measure how much these represent features of presumed relevance. The filtered features are then processed in rectification and pooling layers, which transform data and reduce feature dimensionality for further analysis. The procedure can be replicated along stacked layers, resulting in a hierarchy of increasingly complex features. The final processing layer is a fully connected network that resembles a conventional ANN and is where classification outputs are generated.

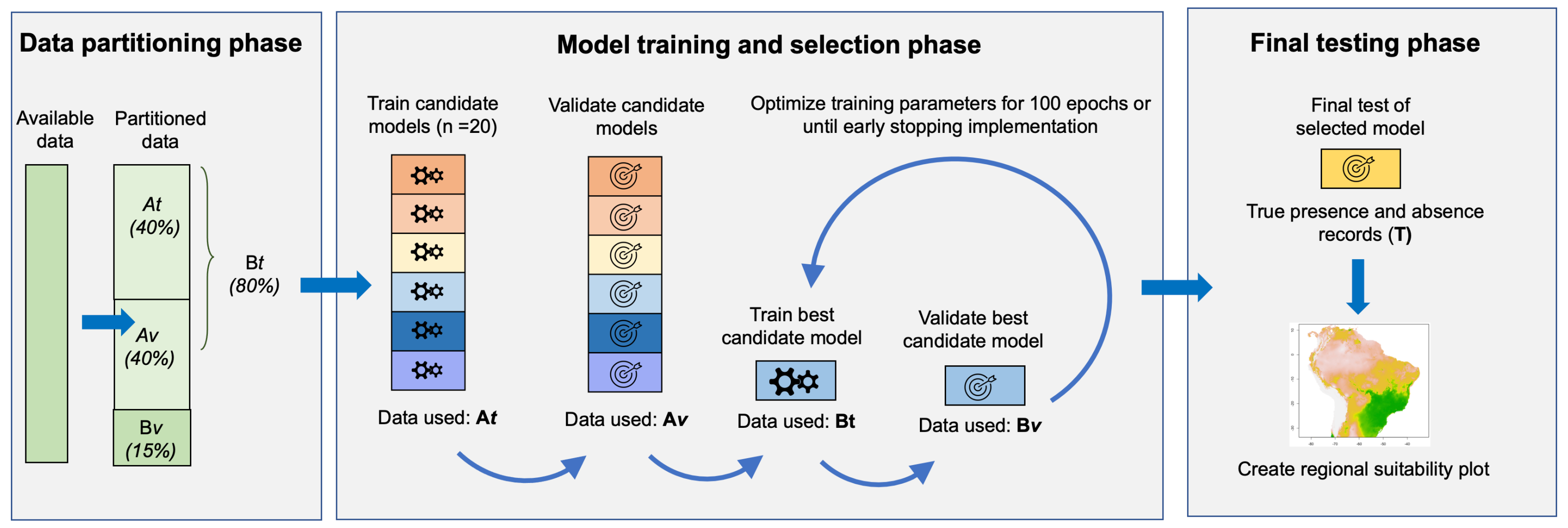
Data partitioning phase
Available data
Partitioned data
At (40%)
Av (40%)
Bv (15%)
Bt (80%)
Model training and selection phase
Train candidate models (n =20)
Validate candidate models
Optimize training parameters for 100 epochs or until early stopping implementation
Train best candidate model
Validate best candidate model
Data used: At
Data used: Av
Data used: Bt
Data used: Bv
Final testing phase
Final test of selected model
True presence and absence records (T)
Create regional suitability plot

Figure 3: Schematic of data partitioning and modeling protocol**.**

RESULTS:

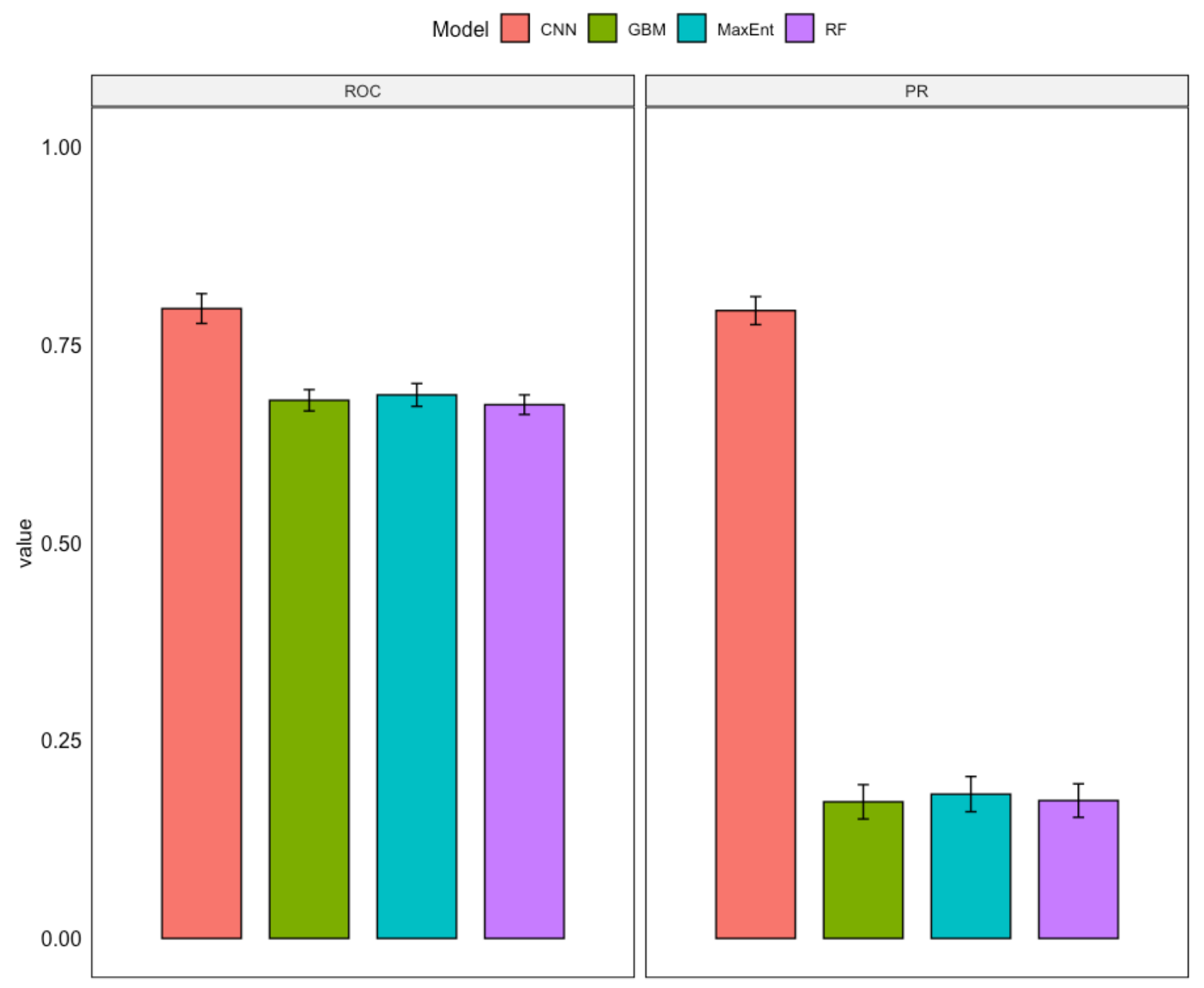

Figure 4: Comparison of average ROC and PR scores for CNNs and conventional methods. CNN models routinely scored higher values, especially when measuring PR curves.

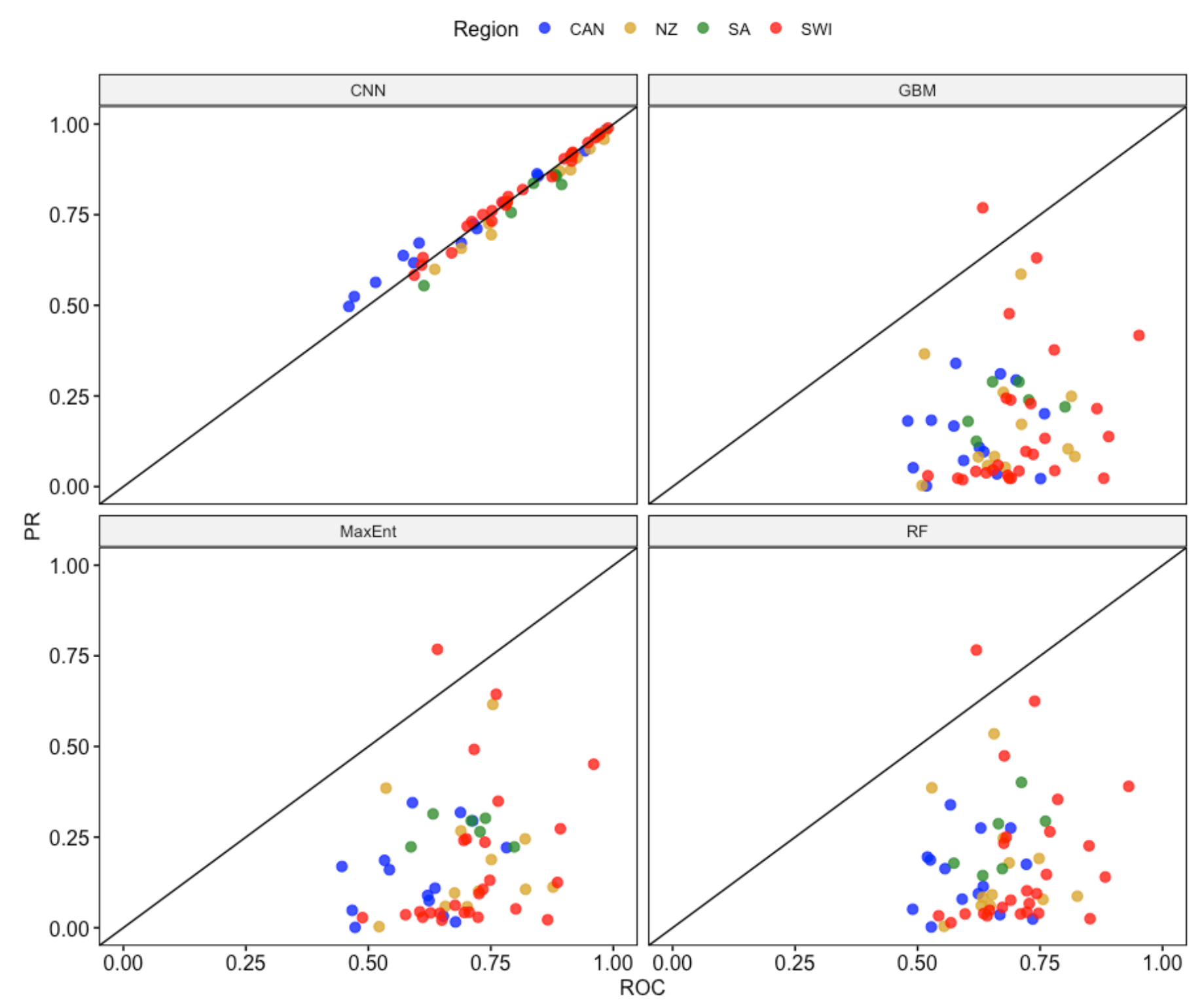

Figure 5: Comparison between ROC and PR scores for each species across different algorithms. CNN models shared a strong linear relationship between statistics as opposed to the conventional methods.

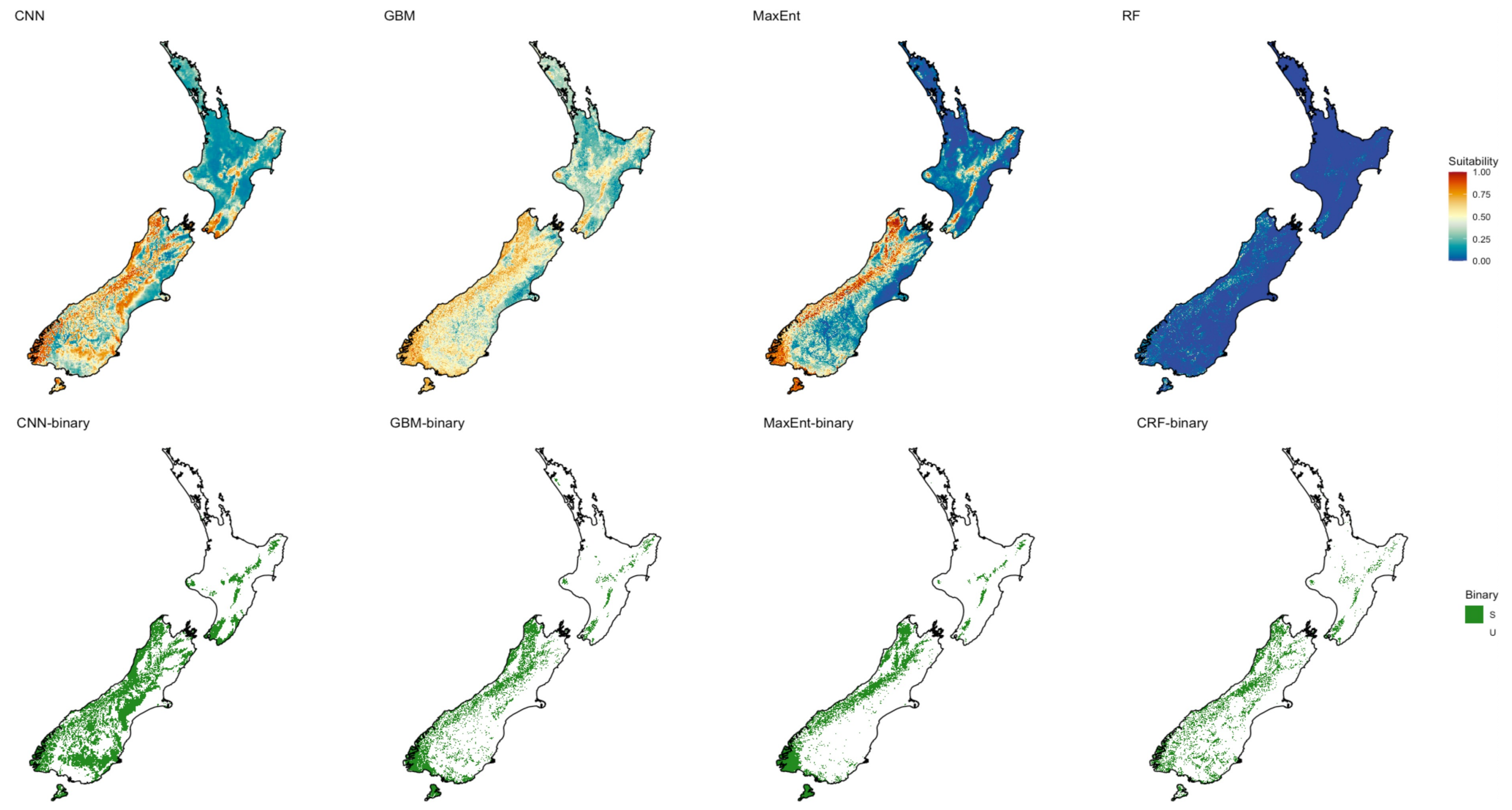

Figure 6: Maps of predicted likelihood of occurrence (relative suitability) and converted binary ranges for species nz08 in the New Zealand data. CNN produced the best results on presence-absence records (ROC = 0.93; PR = 0.91), followed by MaxEnt (ROC = 0.75; PR = 0.19), GBM (ROC = 0.71; PR = 0.17), and RF (ROC = 0.70; PR = 0.18).

WORK CITED:

- Capinha, C., Ceia-Hasse, A., Kramer, A. M., & Meijer, C. (2021). Deep learning for supervised classification of temporal data in ecology. *Ecological Informatics*, *61*, 101252.
- Ceia-Hasse, A., Sousa, C. A., Gouveia, B. R., & Capinha, C. (2023). Forecasting the abundance of disease vectors with deep learning. *Ecological Informatics*, *78*, 102272.
- Elith, J., Graham, C., Valavi, R., Abegg, M., Bruce, C., Ferrier, S., ... & Zimmermann, N. E. (2020). Presence-only and presence-absence data for comparing species distribution modeling methods. *Biodiversity informatics*, *15*(2), 69-80.
- Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International journal of climatology*, *37*(12), 4302-4315.
- He, X., Zhao, K., & Chu, X. (2021). AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems*, *212*, 106622.
- Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., ... & Ovaskainen, O. (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological monographs*, *89*(3), e01370.
- Smith, A. M., Capinha, C., & Kramer, A. M. (2022). Predicting species distributions with environmental time series data and deep learning. *bioRxiv*, 2022-10.
- Sofaer, H. R., Hoeting, J. A., & Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, *10*(4), 565-577.
- Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J., & Elith, J. (2022). Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecological Monographs*, *92*(1), e01486.
- Van Kuppevelt, D., Meijer, C., Huber, F., van der Ploeg, A., Georgievska, S., & van Hees, V. T. (2020). Mcfly: Automated deep learning on time series. *SoftwareX*, *12*, 100548.

Main finding:

This study found both conceptual and practical benefits to using time series data with deep learning algorithms for species distribution modeling.