Developing a Deep Learning Framework for Tree Species Frequency Modelling Across North America

by

Zach Zimmerman

A capping research project report submitted in partial fulfillment of the requirements for the degrees of:

Master of Forestry, University of Alberta, Edmonton, Alberta

and

Master of Science in Environmental Forestry, Bangor University, Wales

Abstract

Accurately representing the spatial distribution and relative abundance of tree species is fundamental to forest inventory, regeneration planning, and climate-informed management. A wide range of species distribution models and spatial forest inventory products already exist, but leading inventory-based products are developed within national boundaries. Because climate-driven shifts in suitable habitat are not constrained by political borders, there is value in modelling frameworks that integrate forest inventory data consistently across jurisdictions, while incorporating climate, topographic, and land-cover information at continental extents.

In this study, I develop a deep learning framework to model tree species frequencies across North America by integrating forest inventory and ecological plot data with environmental predictors. Forest inventory data from the United States, Canada, and Mexico were harmonized to produce proportional species frequency estimates, which were paired with historical climate normals (1951–1980), derived topographic indices, and probabilistic land-cover predictions generated using a separate deep neural network trained remotely sensed land-cover data.

To address the zero-inflated nature of species frequency data, a two-stage modelling approach was implemented, consisting of a presence–absence classifier followed by a conditional frequency regression model. Additional preprocessing steps, including spatial aggregation of plots, filtering of observations using buffered historical species ranges, and the introduction of pseudo-plots in non-forested regions, were applied to improve computational efficiency and ecological realism.

Model performance was evaluated using withheld inventory data and spatial comparisons with historical species range maps for a regionally diverse subset of tree species. Results show that incorporating topographic variables and probabilistic land-cover information improves model performance relative to climate-only formulations, and produces spatially coherent, ecologically plausible species frequency patterns across different forest regions of the continent.

The framework presented here provides a foundation for generating consistent, continent-wide species frequency surfaces that complement existing forest inventory products and support applications in forest inventory, regeneration planning, and future climate-informed analyses.

Acknowledgements

I would like to express my deepest gratitude to Dr. Andreas Hamann for his invaluable support and guidance throughout this project. His mentorship, constructive feedback, and steady encouragement were instrumental in shaping the direction and quality of this research. I am especially thankful for the many discussions, reviews, and insights he provided during every stage of the process.

I also wish to thank Nicholas Boyce, my collaborator in the SIS lab, for his thoughtful feedback and for the development of the collaborating landcover model. His expertise and collaborative spirit greatly enriched this work.

Table of Contents

1. Introduction

Forests occupy approximately 30% of the Earth's land surface and play a central role in supporting terrestrial biodiversity, regulating climate, and providing ecosystem services to human societies. They function as major carbon sinks (Pan et al. 2011), influence energy and water exchanges between the land surface and atmosphere (Chapin et al. 2008), and underpin economic and cultural values associated with forestry and land management. As climate conditions continue to change, the ability to represent where tree species occur, and in what relative abundance, has become increasingly important for forest inventory, regeneration planning, and climate-informed decision-making (Booth 2018; Dar et al. 2022; Esquivel-Muelbert et al. 2019; Massey et al. 2023).

A wide range of spatial forest inventory products has been developed to extend plot-based measurements across landscapes using statistical imputation and environmental similarity approaches. In Canada, the National Forest Inventory has been used to generate continuous maps of forest attributes, including tree species composition and relative abundance, at approximately 250 m resolution by linking ground plots with climate, topographic, and remotely sensed predictors using k-nearest neighbor methods (Beaudoin et al. 2014; National Forest Inventory 2024). In the United States, comparable nearest-neighbor and gradient nearest-neighbor imputation approaches have been developed using Forest Inventory and Analysis (FIA) plots to produce spatially explicit maps of forest composition at similar spatial resolutions (Ohmann et al. 2011; Wilson et al. 2012). These products provide robust and widely used representations of forest composition within their respective jurisdictions and form an important foundation for forest monitoring and management.

Because these inventory-based products are developed independently within national boundaries, they necessarily represent continental species distributions only in part when species ranges extend across borders. This reflects differences in inventory design, data availability, and modelling frameworks across countries, rather than limitations of the underlying methods. Complementary to these existing efforts, there is value in developing a modeling framework that integrates forest inventory data consistently across jurisdictions, while incorporating environmental predictors known to influence species distributions. Climate, topography, and land cover jointly shape where tree species can occur and how abundant they are within forested landscapes (Lee-Yaw et al. 2022). Because climatic conditions and associated habitat changes operate across biogeographic space rather than political boundaries, applications such as regeneration planning, seed transfer, and assisted migration under climate change benefit from continuous, continental-scale representations of species frequencies.

Machine learning methods have become increasingly prominent in species distribution modelling due to their ability to capture nonlinear relationships and interactions among environmental predictors (Evans et al. 2011). Among these approaches, deep neural networks offer advantages for large-scale applications, including scalability to large datasets, flexibility in handling high-dimensional predictors, and strong predictive performance when sufficient training data are available (Botella et al. 2018; LeCun et al. 2015; Valavi et al. 2022). Rather than replacing established modelling approaches, deep learning provides a complementary set of tools that can be integrated into multi-stage workflows linking environmental data, forest inventories, and spatial prediction.

In this study, I develop a continental-scale deep learning framework to model tree species frequencies across North America by integrating forest inventory and ecological plot data with climate, topographic, and land-cover information. Forest inventory datasets from the United States, Canada, and Mexico are harmonized to produce proportional species frequency estimates, which are paired with historical climate normals, derived topographic indices, and probabilistic land-cover predictions generated using a separate deep neural network. A two-stage modelling approach is employed to address the zero-inflated nature of species frequency data, separating the prediction of species occurrence from the prediction of relative abundance where species are present (Martin et al. 2005; Rozenberg 2010).

1. 1. Objectives

The primary objective of this work is to develop a flexible and scalable framework for generating consistent, continent-wide species frequency surfaces. These outputs are intended to complement existing inventory-based products and support applications in forest inventory, regeneration planning, and future climate-informed analyses.

Specifically, this thesis addresses the following objectives:

1. Harmonize forest inventory and ecological plot data across national datasets from the United States, Canada, and Mexico, and derive consistent proportional measures of tree species frequency suitable for continental-scale modeling.

2. Develop a two-stage deep learning modeling framework that separately predicts species occurrence and relative abundance, addressing the zero-inflated nature of forest inventory data and enabling spatially explicit frequency mapping.

3. Evaluate the contribution of climate, topographic and probabilistic land-cover predictors to species frequency modeling, and assess how the inclusion of these variables influences model performance relative to climate-only formulations.

4. Generate and assess continent-wide species frequency maps for a representative subset of tree species, evaluating spatial coherence and ecological plausibility through comparisons with withheld inventory data and historical species range maps.

2. Methods

2.1. Forest inventory and ecological plot data

To develop species distribution models (SDMs) across North America, I compiled a harmonized dataset of georeferenced forest inventory and ecological plot data from national and regional databases spanning the United States, Canada, and Mexico (Fig. 1). The dataset includes both permanent and temporary plots with species presence and abundance information, linked to standardized geographic coordinates for model training and evaluation.

For the United States, plot data were obtained from the Forest Inventory and Analysis (FIA) database managed by the U.S. Forest Service (U. S. Forest Service 2023). The FIA program conducts continuous forest monitoring across all states and provides tree- and plot-level measurements of species composition, tree height, and diameter.

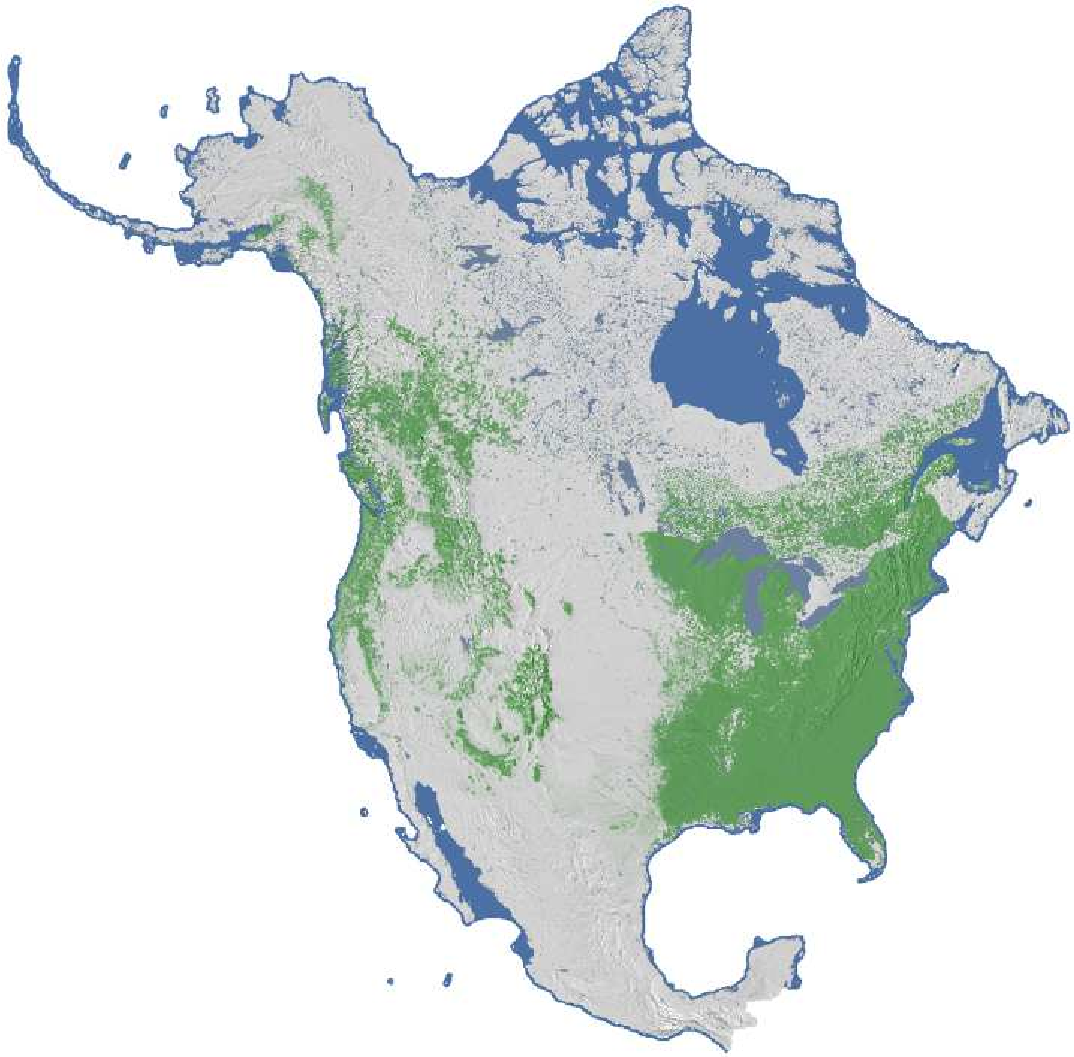

Fig. 1 Forest inventory plots of U.S. and Canada

In Canada, plot data were sourced from several complementary initiatives. The National Forest Inventory (NFI), coordinated by Natural Resources Canada, provides a systematic, grid-based sampling of forest conditions across the country (NRCAN 2021). Additional provincial-level data were obtained from the Multi-Agency Ground Plot (MAGPlot) database, which aggregates forest ground plot data from provincial forestry agencies into a harmonized, Canada-wide system

(National Forest Inventory 2024). MAGPlot includes standardized records of tree species, diameter, crown attributes, and site conditions.

Additionally, provincial ecological plot databases were used where available. For Alberta, data from the Ecological Site Information System (ESIS) were included, providing detailed vegetation and ecological site descriptions with percent canopy cover projected to the ground for multiple canopy layers (Alberta Environment & Parks 2021). In British Columbia, plot data supporting the Biogeoclimatic Ecosystem Classification (BEC) system were used; these data similarly record percent canopy cover projected to the ground across multiple canopy layers (Meidinger and Pojar 1991).

For Mexico, plot data were incorporated from the Inventario Nacional Forestal y de Suelos maintained by the National Forestry Commission (Comisión Nacional Forestal 2020). This database includes systematic measurements of forest composition, structure, and biomass across major vegetation zones, compiled over multiple inventory cycles. Tree-level measurements include height and diameter, and all available inventory cycles were used in this study.

2.2. Predictor variables

2.2.1. Climatic predictors

Climate predictor variables were generated with the ClimateNA software package (Wang et al., 2016), which provides high-resolution climate surfaces for North America based on interpolated weather station data, digital elevation models, and environmental lapse rate based downscaling. Here, we use historical climate normals for the 1951–1980 period, representing climate conditions that largely predates anthropogenic climate warming, while still being represented with a good weather station network to infer high resolution climate grids (Fig. 2). A period that predates major anthropogenic climate warming was chosen as a baseline to satisfy the assumption that tree species distributions are in approximate equilibrium with those climate conditions.

A total of 16 climate variables representing temperature, precipitation, and moisture balance were included as predictors. These comprised mean annual temperature (MAT), mean warmest month temperature (MWMT), mean coldest month temperature (MCMT), temperature difference between MWMT and MCMT (TD), mean annual precipitation (MAP), mean summer precipitation (MSP), degree days below 0 °C (DD0), degree days above 5 °C (DD5), precipitation as snow (PAS), extreme minimum temperature over a 30-year period (EMT), Hargreaves' climatic moisture deficit (CMD), mean annual relative humidity (RH), annual heat–moisture index (AHM), summer heat–moisture index (SHM), and Hogg's climate moisture index (CMI).

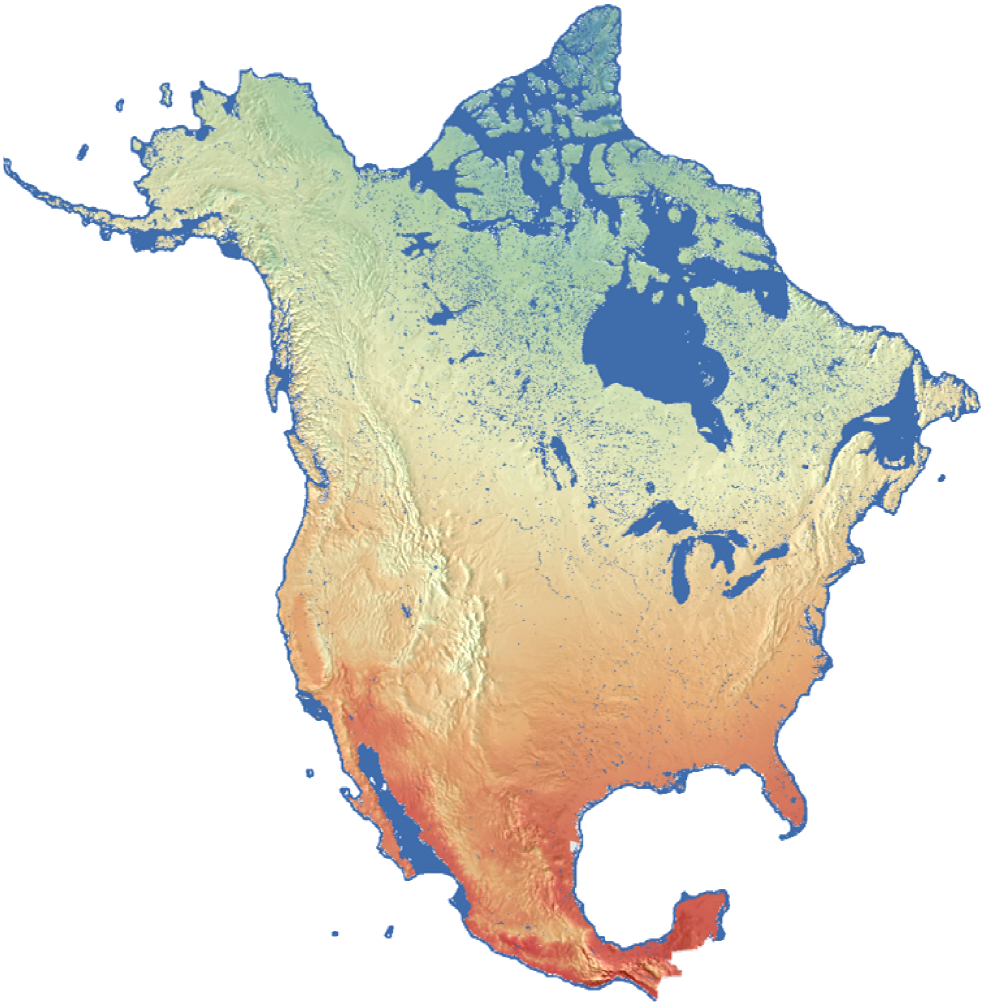

Fig. 2. Mean annual temperature of North America from ClimateNA

Prior to model training, all climate variables were standardized to a mean of zero and unit variance so that each predictor entered the model on a comparable scale, allowing the network to initially treat all variables as having potentially equal influence during training. Temperature-based variables (MAT, MWMT, MCMT, TD, EMT, DD0, DD5, RH) were standardized without prior transformation. In contrast, precipitation- and moisture-related variables (MAP, MSP, PAS, CMD, AHM, SHM, CMI) exhibited strong positive skew and were log-transformed prior to standardization. A generalized log(x + k) transformation was applied, where k is a variable-specific constant selected to adjust the strength of the transformation and improve approximation to normality. The value of k was chosen individually for each variable, with negative values permitted for stronger transformations provided that all transformed values remained within the domain of the logarithmic function. This approach reduced skewness, stabilized variance, and improved numerical behavior during neural network training.

2.2.2. Topographic predictors

Additional predictor variables describing landscape structure were included to improve model accuracy and better characterize habitat suitability. Topographic predictors were derived from the MERIT digital elevation model, resampled to a 1 km spatial resolution to match the target resolution of the species frequency predictions and climate predictor grids.

A total of 14 topographic variables were generated to capture terrain position, exposure, and the influence of major water bodies. These variables included measures of terrain exposure, weighted by prevailing wind direction (see details below) and calculated at multiple spatial scales (1km, 2km and 4km), hillshade calculated under both south-facing solar angle, distance to lakes weighted by prevailing wind direction with maximum distances of 100 and 250 km, distance to ocean weighted by prevailing wind direction with maximum distances 1000 and 2500 km, a Compound Topographic Index (CTI) calculated at scales of 1km, 2km and 4km (Fig. 3), and the Topographic Position Index (TPI) calculated at the same three spatial scales.

Topographic Position Index (TPI) describes the relative position of a location within the surrounding terrain, distinguishing ridge tops (positive values), valley bottoms (negative values), and flat or mid-slope positions (values near zero), independent of absolute elevation. The

Compound Topographic Index (CTI), also referred to as a topographic wetness index, estimates the potential for water accumulation based on upslope contributing area and local slope, and provides an index of site-level moisture availability relevant to vegetation patterns.

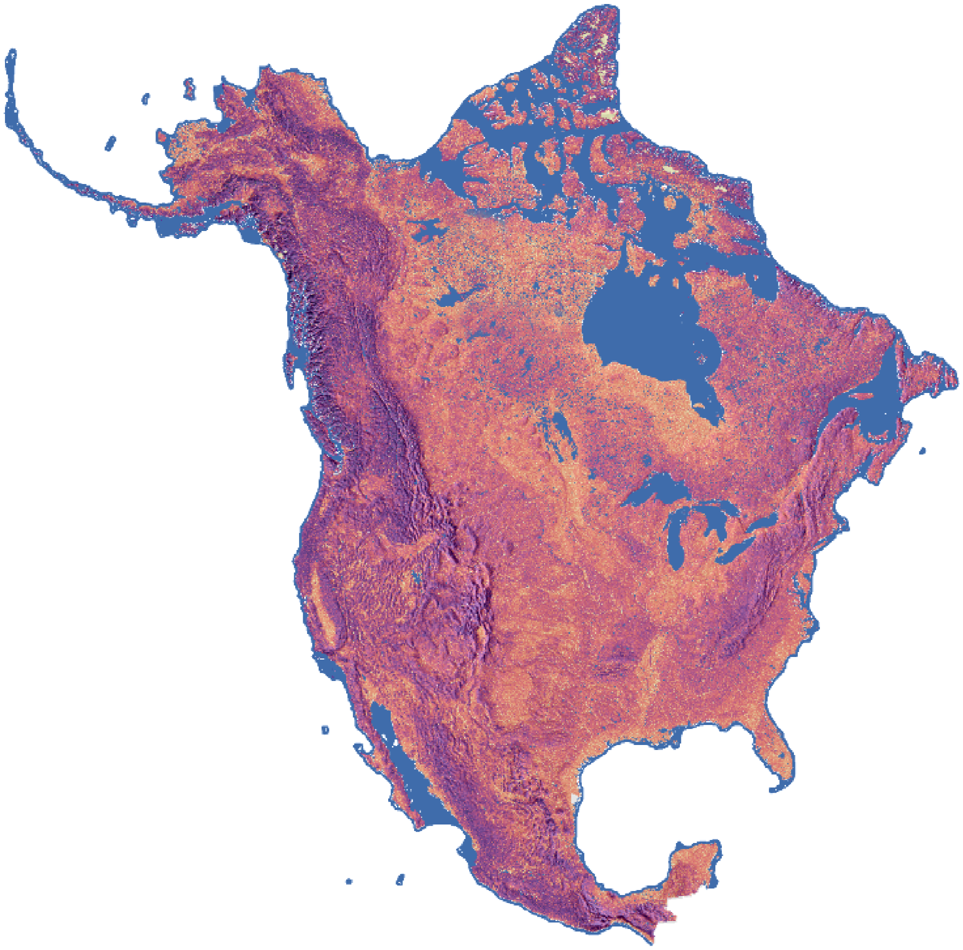

Fig. 3. Compound Topographic Index with landscape level resolution of North America.

To account for the fact that topographic influences on climate and atmospheric processes act across multiple spatial scales, several topographic predictors were generated at different resolutions and spatial extents. These multi-scale representations allow the modelling framework to evaluate the relative importance of fine-scale versus broader-scale topographic controls on species frequency during training. All topographic variables were transformed and scaled prior to modelling to ensure comparable numerical ranges and stable neural network training, following the same general preprocessing principles applied to climate predictors.

2.2.3. Land cover probabilities

As additional predictor variables, I incorporated probabilistic estimates of land cover derived from a predictive land cover model developed by Boyce (2025, unpublished). These predictors represent the likelihood of different natural land cover types based on climate and topographic conditions, rather than observed contemporary land use. This distinction allows species frequency models to reflect climatically and physiographically suitable habitat while drawing on remotely sensed land cover data with complete continental coverage, thereby providing a substantially broader spatial foundation than plot-based observations alone, except in areas where land cover has been altered by human conversion.

The land cover model was trained using MODIS land cover classification data, in which land cover classes served as the dependent variable. Agricultural and urban classes were excluded from the training process, but their spatial locations were retained as prediction targets. Predictor variables included a subset of the climate and topographic variables described above, capturing broad-scale climatic gradients as well as terrain position, exposure, and proximity to water bodies. A deep neural network classifier was trained to predict the probability of natural land cover classes for each grid cell, following general methodological principles described in previous work (Namiiro et al. 2025; Boyce (2025, unpublished**)**.

After training, the land cover model was applied across the study area using the same set of predictor variables. For grid cells currently classified as agriculture or urban, the model predicted the most probable natural land cover class based on climate and topography (Fig. 4).

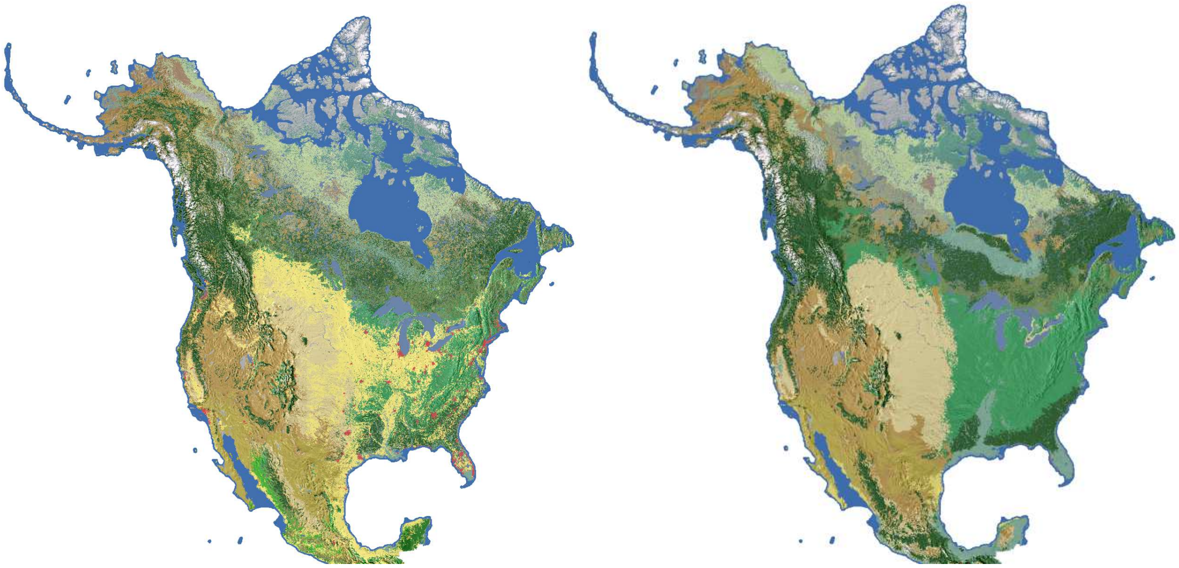

Fig. 4 MODIS remote sensed land cover classes (left) backfilled land cover classifications with urban and agricultural land removed from training and output (right).

Rather than using a single categorical land cover assignment, the resulting class probabilities were retained and used as continuous predictor variables in the tree species frequency models. For example, the predicted probability of deciduous forest cover (Fig. 5) was included as an input variable for modelling species associated with deciduous-dominated ecosystems.

Incorporating probabilistic land cover predictors allows the species frequency models to condition predictions on the likelihood of forest cover under natural conditions, improving ecological realism in regions where contemporary land use obscures climatic suitability for tree species. This approach supports applications focused on climate suitability, seed sourcing, and regeneration planning, where potential habitat is of greater relevance than realized land use. However, because the resulting species frequency surfaces represent climatically suitable forest cover rather than current land cover, they must be combined with observed remotely sensed land cover data when used for applications requiring representation of present-day forest extent or land use.

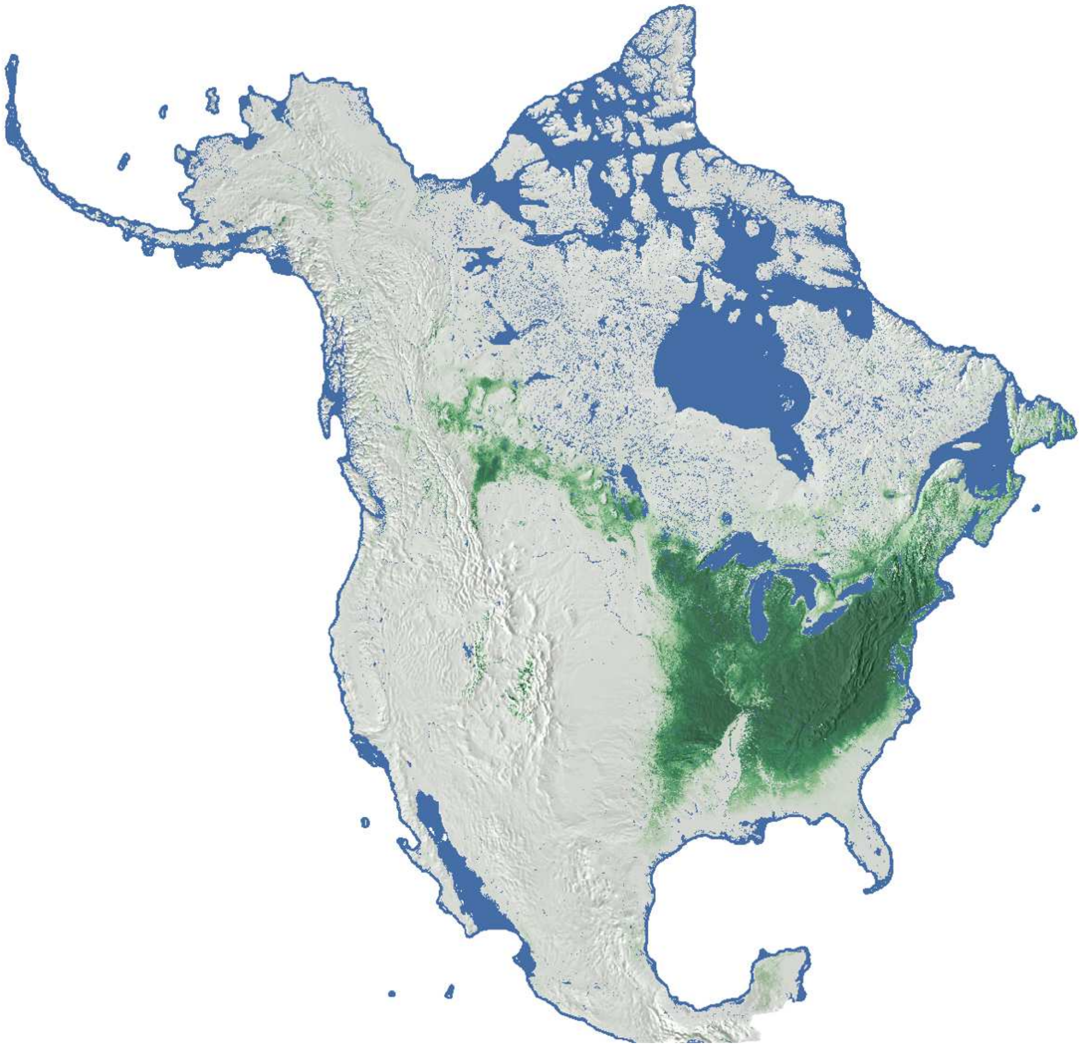

Fig. 5 Probability of deciduous forest as land cover class across North America

2.3 Forest inventory harmonization and training data preparation

To ensure consistency across heterogeneous forest inventory datasets and to provide an ecologically meaningful response variable for modeling, a series of harmonization, scaling, and data preparation steps were applied prior to model training. These steps define how species frequency was quantified and address known sources of bias and imbalance in large-scale forest inventory data.

2.3.1 Plot data harmonization and definition of species frequency

Species frequency was quantified using proportional measures of species dominance derived from forest inventory plot data. Depending on data availability and measurement protocols of each source dataset, species dominance was calculated as either the percentage of basal area or the percentage of crown cover projected to the ground. This approach allowed the use of multiple national and regional inventory datasets while maintaining a consistent response variable suitable for modelling. When multiple measurements were available for a given plot, either from different canopy layers or from repeated sampling events, species-level values were pooled and rescaled so that total tree species frequency summed to 100% within each plot. To further standardize species frequency estimates across ecosystems, proportional basal area or canopy cover values were scaled using modeled land cover class probabilities as described above.

Specifically, predicted probabilities of forest land cover types were used to scale tree species frequencies such that the sum of forest species frequencies plus the proportion of non-forested land equalled 100% of the ecosystem land base. Because agricultural and urban land cover classes were excluded from the training of the land cover model and subsequently reclassified to the most probable natural land cover type, this procedure yields proportional estimates of forest tree species frequencies expected under undisturbed conditions. The resulting response variable therefore represents relative species frequency within climatically and physiographically suitable forest habitat, rather than realized contemporary land use.

2.3.2 Plot Aggregation

The modelling dataset initially comprised approximately 800,000 individual forest inventory plots distributed across North America. To reduce potential bias arising from highly uneven sampling intensity among regions and inventories, plot aggregation was applied prior to model training. Without aggregation, areas with dense inventory coverage could disproportionately influence model fitting, potentially overshadowing rarer but ecologically important observations from sparsely sampled regions.

Plots were aggregated based on spatial proximity, elevation similarity, and ecosystem context. Specifically, only plots falling within defined horizontal distance and elevation windows and within the same Level-4 ecoregion (ecosystem variant) were eligible for aggregation. This constraint ensured that aggregation occurred only among plots sharing comparable ecological settings. Predictor variables were first extracted at the original plot locations and subsequently aggregated alongside species frequency data, so that environmental information reflected the same spatial, elevational, and ecological context as the response variable. Mean species frequencies and predictor values were calculated for each aggregated unit.

Multiple aggregation schemes were evaluated to assess their influence on model behaviour and ecological realism. No meaningful degradation in predictive performance was observed under increasingly coarse aggregation, indicating that dominant species–environment relationships were preserved. Based on these tests, a final aggregation threshold of 50 km horizontal distance and 250 m elevation difference within the same Level-4 ecoregion was selected, reducing the dataset from approximately 800,000 individual plots to approximately 29,000 aggregated observations while mitigating sampling-density bias across the continent.

2.3.3 Filtering for Potential Misidentification

To reduce the influence of potential species misidentifications, recent introductions, or non-native occurrences on model training, species occurrence data were filtered using digitized historical species range maps published by (Little 1971). Three filtering strategies were evaluated: (1) no filtering, in which all observations were retained; (2) strict filtering, in which all plot records falling outside the historical range of a species were removed; and (3) buffered filtering, in which a 200 km buffer was applied around each species' historical range to account for spatial uncertainty in the range maps and limited natural dispersal beyond recorded boundaries. For each strategy, models were retrained and evaluated to assess impacts on predictive performance and ecological realism. Buffered filtering provided the best balance between retaining sufficient training data and excluding ecologically implausible occurrences. Consequently, observations falling outside the 200 km buffered historical range were excluded from model training (Fig. 6, Douglas-fir example).

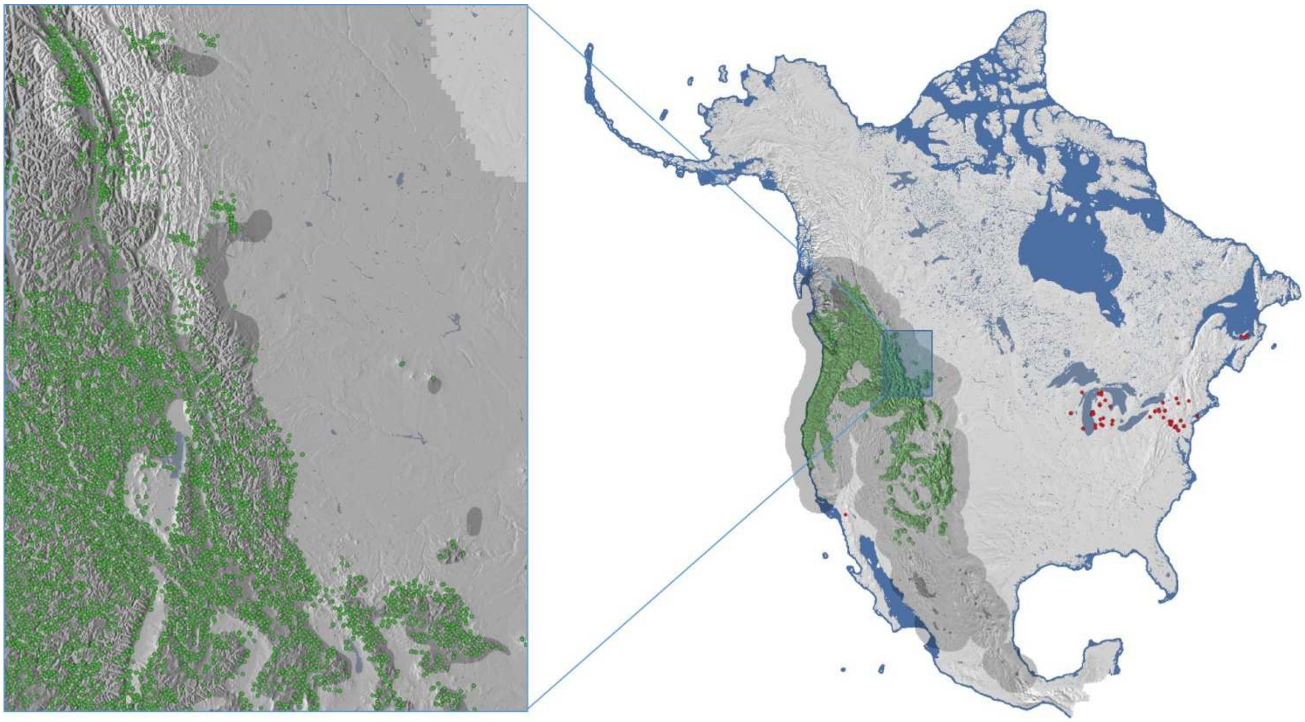

Fig. 6 Map of historic range, range expanded by 200km, and observations of Douglas fir, with red plots being excluded

2.3.4 Introduction of pseudo-plots in non-forested regions

Forest inventory data are inherently limited to forested regions and therefore lack observations in treeless environments such as tundra, deserts, and alpine areas. Early model iterations trained exclusively on inventory plots exhibited a tendency to overpredict species presence in such regions. To address this bias, pseudo-plots representing true absences were introduced into non-forested areas. Non-forested regions were identified using remotely sensed MODIS vegetation cover data in combination with probabilistic land cover predictions. Grid cells were classified as treeless if they exhibited no observed tree or shrub cover in MODIS data and had a predicted probability of tree or shrub cover below 5% based on the land cover model. Within these regions, 100,000 pseudo-plots were generated and assigned a species frequency of zero for all tree species. The inclusion of pseudo-plots provided explicit training information on unsuitable habitat and improved the model’s ability to discriminate between forested and non-forested environments across the continent (Fig. 7). This step substantially reduced spatial overprediction and improved ecological plausibility of species frequency maps.

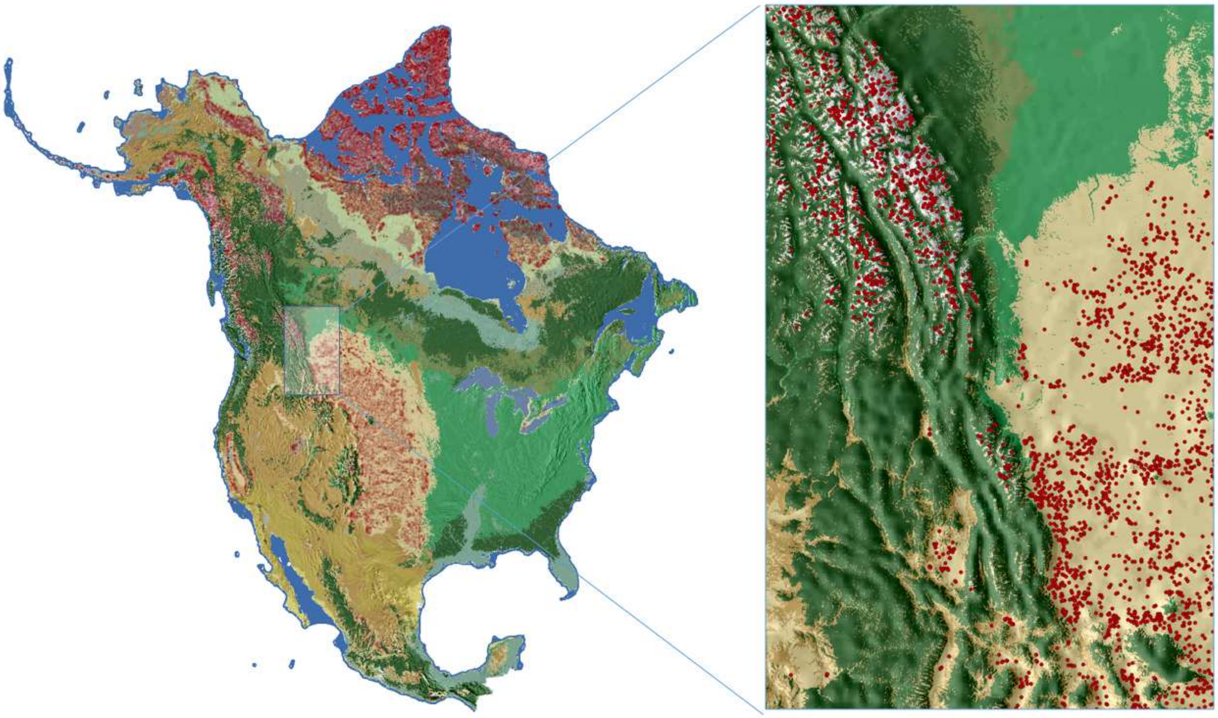

Fig. 7 Introduced pseudo-plots in non-forested areas.

2.4. Modeling framework

Tree species frequency across North America was modelled using a deep neural network (DNN) framework that integrates climate, topographic, and land cover predictors with harmonized forest inventory data. The modeling workflow is summarized in Fig. 8, which illustrates the sequential steps from data preparation to spatial prediction.

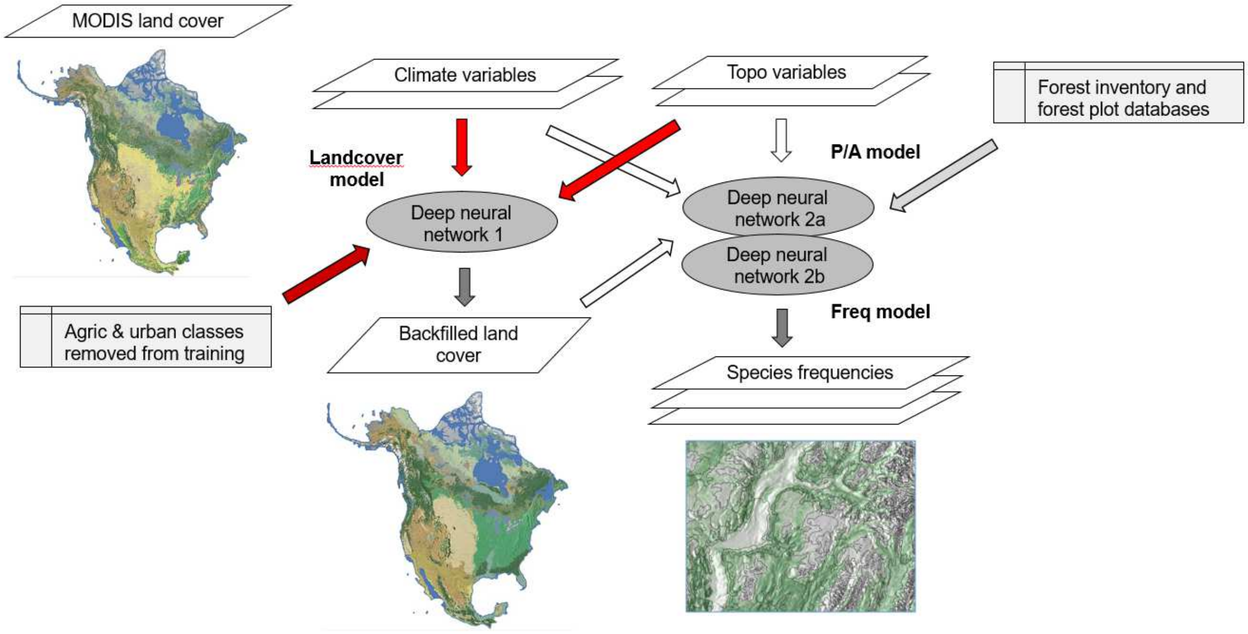

Fig. 8 Overview of the modeling framework

Aggregated forest inventory plot data were paired with corresponding predictor variables, including interpolated climate data from ClimateNA, topographic metrics derived from digital elevation models, and probabilistic land cover estimates. These data formed the training dataset used to develop species-specific models capable of predicting relative species frequency across continuous geographic space.

To address the strongly zero-inflated nature of species frequency data, a two-part modeling strategy was implemented using a hurdle-model framework (Martin et al. 2005; Rozenberg 2010). Zero values arise both from true absences and from limited sampling coverage, and modeling occurrence and abundance as a single process can lead to biased predictions. The hurdle framework separates these processes into two sequential components, allowing presence and relative abundance to be modeled independently.

For each species, two neural networks were trained. The first network (Fig. 8, DNN 2a) was a binary classifier predicting species presence or absence (frequency > 0). This model used a feedforward architecture with a single hidden layer (64 rectified linear units, ReLU), dropout regularization, and a sigmoid output layer. The model was trained using binary cross-entropy loss. The second network (Fig. 5, DNN 2b) was a conditional regression model trained only on plots where the species was present. This network predicted relative species frequency using a deeper feedforward architecture with three hidden layers (256, 128, and 64 units), batch normalization, dropout regularization, and a linear output layer. The log-cosh loss function was used to reduce sensitivity to outliers while retaining stability during optimization.

After training, predictions from the presence and conditional frequency models were combined to generate final species frequency estimates. The probability of presence from the classifier was multiplied by the predicted conditional frequency to yield expected species frequency at each location. This combined output produced spatially continuous, continent-wide predictions of relative species frequency under historical environmental conditions. Model predictions were generated on a regular grid across North America at 1 km spatial resolution using interpolated ClimateNA variables and the corresponding topographic and land cover predictors. The resulting outputs were rasterized to produce continuous species frequency surfaces suitable for spatial analysis and visualization.

2.5. Model evaluation

For each species, the aggregated plot dataset was randomly partitioned into training and testing subsets, with 70% of observations used for model fitting and the remaining 30% withheld for independent evaluation. This split was applied consistently across both components of the hurdle model to ensure comparable evaluation of presence and frequency predictions.

Model performance was assessed using multiple complementary metrics. For the conditional frequency models, predictive accuracy was evaluated using mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (R^2), and bias calculated from predictions on the withheld test data. Presence–absence models were evaluated using standard classification diagnostics, including accuracy and inspection of predicted probability surfaces. These metrics were used to assess both overall model performance and potential systematic bias across environmental gradients.

To evaluate the ecological realism and generality of the modeling framework, additional sensitivity analyses were conducted using a subset of ecologically and economically important tree species representing a range of life histories, climatic tolerances, and geographic distributions. These species included black spruce (*Picea mariana*), Douglas-fir (*Pseudotsuga menziesii*), trembling aspen (*Populus tremuloides*), subalpine fir (*Abies lasiocarpa*), and sugar maple (*Acer saccharum*). These species were selected due to their broad distributions across North America, high representation in forest inventories, and importance to forest management and ecological function.

Model outputs were further evaluated qualitatively by comparing predicted species frequency maps with historical species range maps (Little 1971) and known biogeographic patterns. Visual inspection was used to identify unrealistic spatial extrapolations, national boundary artefacts, or systematic overprediction, particularly near range margins and in climatically marginal regions. These qualitative assessments complemented quantitative metrics and informed refinement of data preparation steps described above.

3. Results & Discussion

3.1. Species distribution maps and ecological plausibility

Predicted species frequency maps showed strong agreement with known biogeographic patterns for the species examined, providing qualitative support for the ecological plausibility of the modelling framework. For western North American species, predicted frequency patterns aligned closely with major climatic gradients, elevation zones, and ecosystem delineations.

Douglas-fir provides an example (Fig. 9). The model captured the broad distribution of coastal Douglas-fir in low-elevation regions of the Pacific Northwest, as well as the more spatially constrained distribution of interior Douglas-fir associated with higher elevations, complex topography, and rain-shadow effects. Predicted frequencies varied systematically across Level 4 ecosystem delineations, reflecting known differences in forest composition and site conditions. Comparison between the 250 m and 1 km resolution maps indicates that spatial aggregation preserved dominant ecological gradients while reducing fine-scale noise.

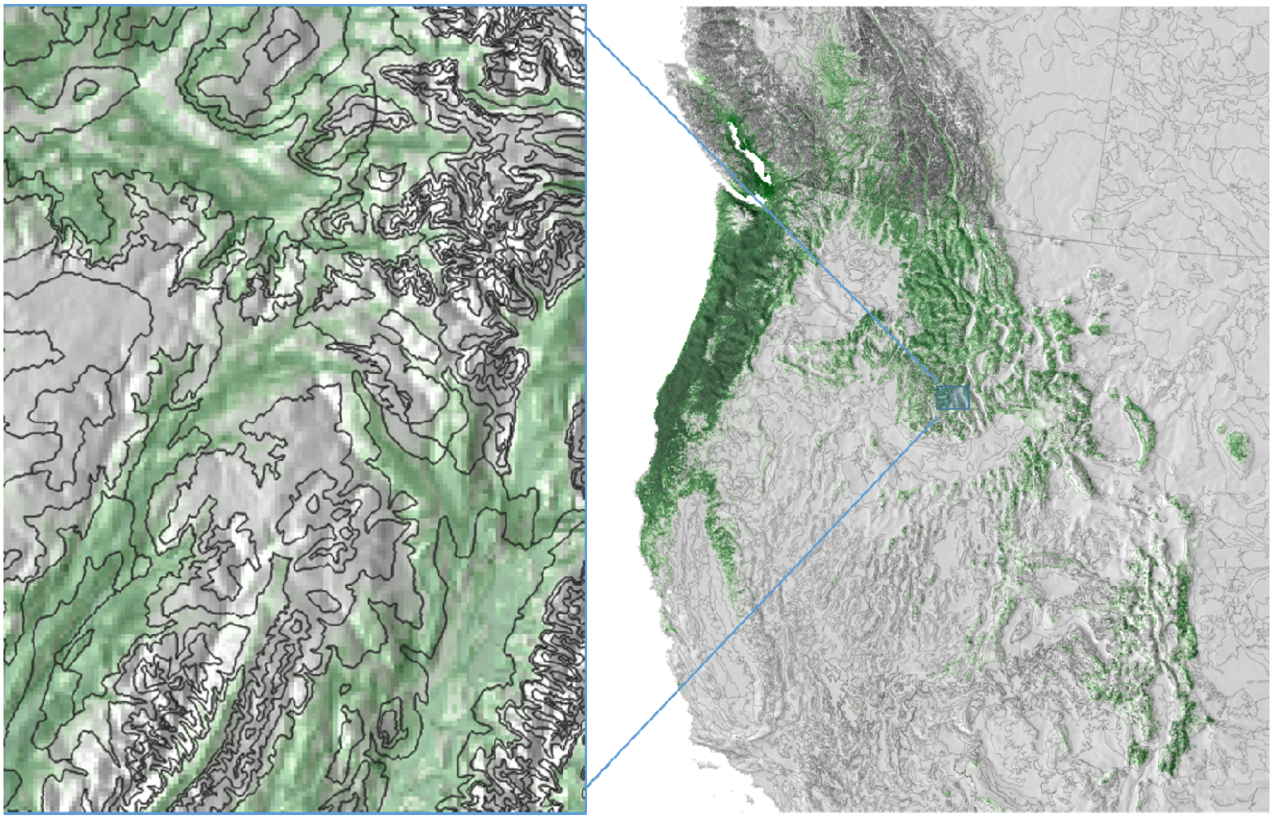

Fig. 9 Species distribution map of Douglas fir at 1 km resolution (right) 250m resolution (left), and level 4 ecosystem delineations.

As another example, black spruce predictions were concentrated in northern and wet boreal environments (Fig. 10). Predicted frequency surfaces showed smooth spatial transitions and respected known range limits, supporting the use of the framework for representing relative species dominance at continental scales.

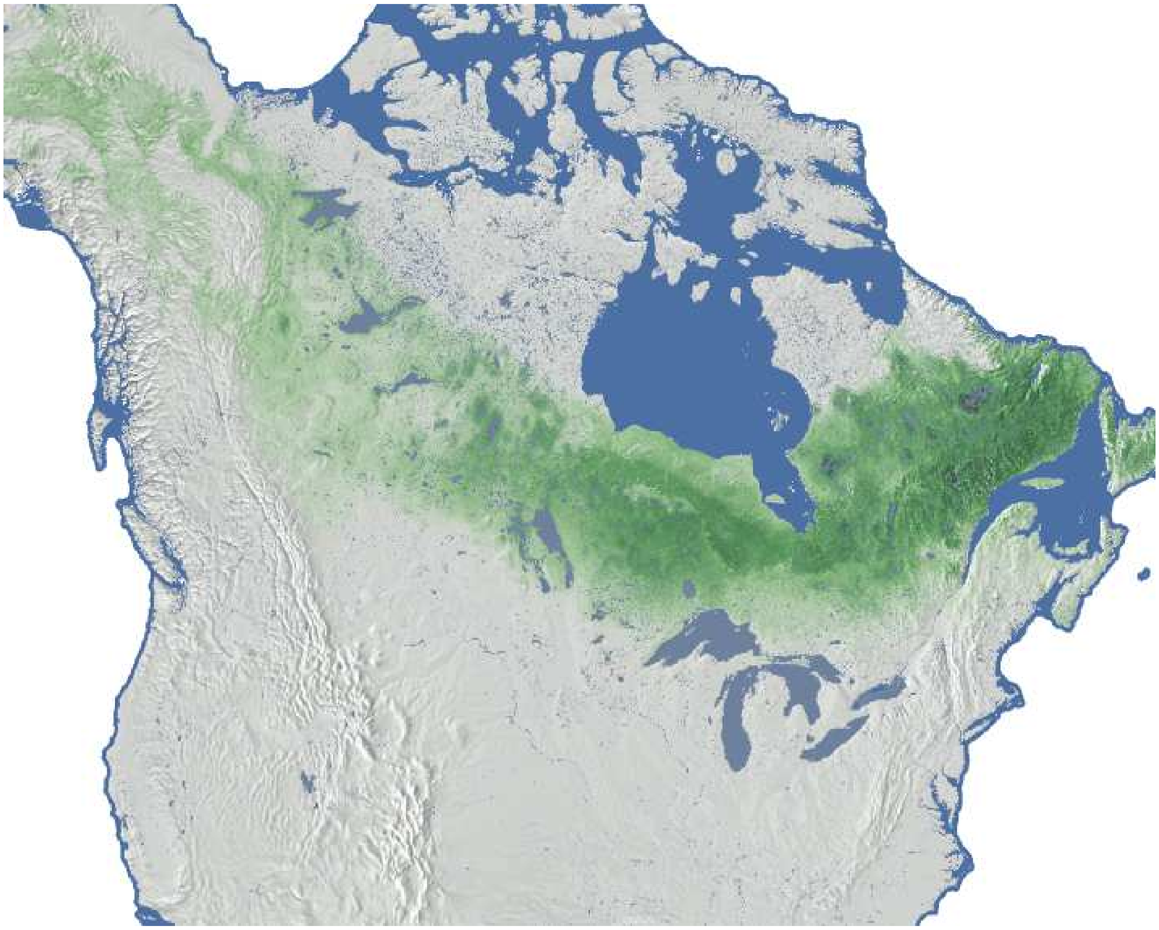

Fig. 10 Species frequency maps of black spruce at 1km Resolution.

3.2. Predictor importance and ecological interpretation

Permutation-based variable importance analyses further support the ecological realism of the models. For Douglas-fir, the presence–absence model (DNN 2a) identified the probability of needle-leaf forest cover as the most influential predictor, followed by climatic constraints related to moisture balance and cold tolerance (Fig. 11). The prominence of variables such as log-transformed annual heat–moisture index and extreme minimum temperature is consistent with established climatic controls on Douglas-fir distribution.

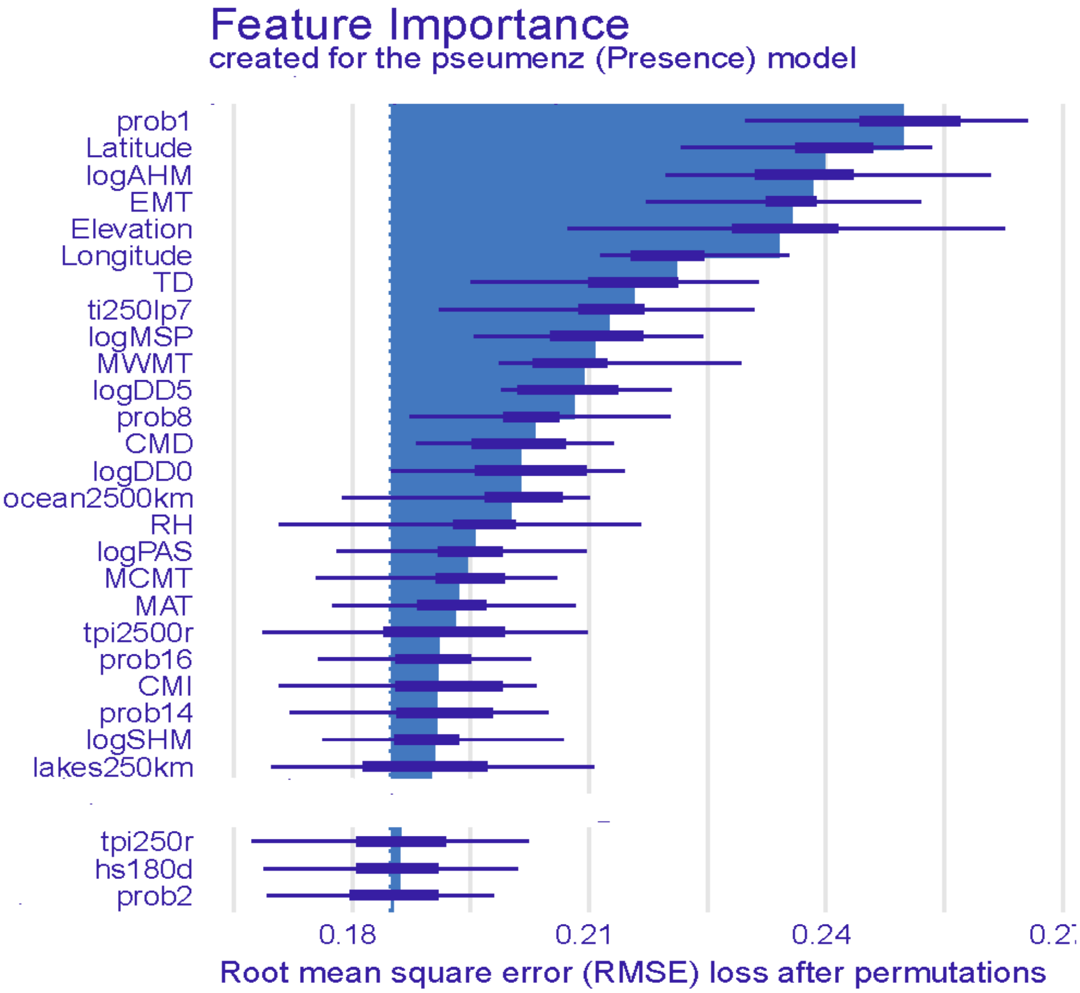

Fig. 11 Variable importance of predictor variables (y-axis) by root mean square error loss after removal of variable for presence absence model (DNN 2a) of Douglas fir.

In contrast, the conditional frequency model (DNN 2b), trained only on plots where Douglas-fir was present, emphasized predictors associated with local abundance rather than broad occurrence limits (Fig. 12). Elevation, topographic orientation, distance to the ocean, and terrain exposure emerged as dominant predictors, with additional contributions from snow-related variables, summer moisture conditions, and shrubland cover probability. This contrast between the two modelling stages illustrates the utility of the hurdle framework for separating factors that govern species occurrence from those influencing relative dominance within suitable habitat.

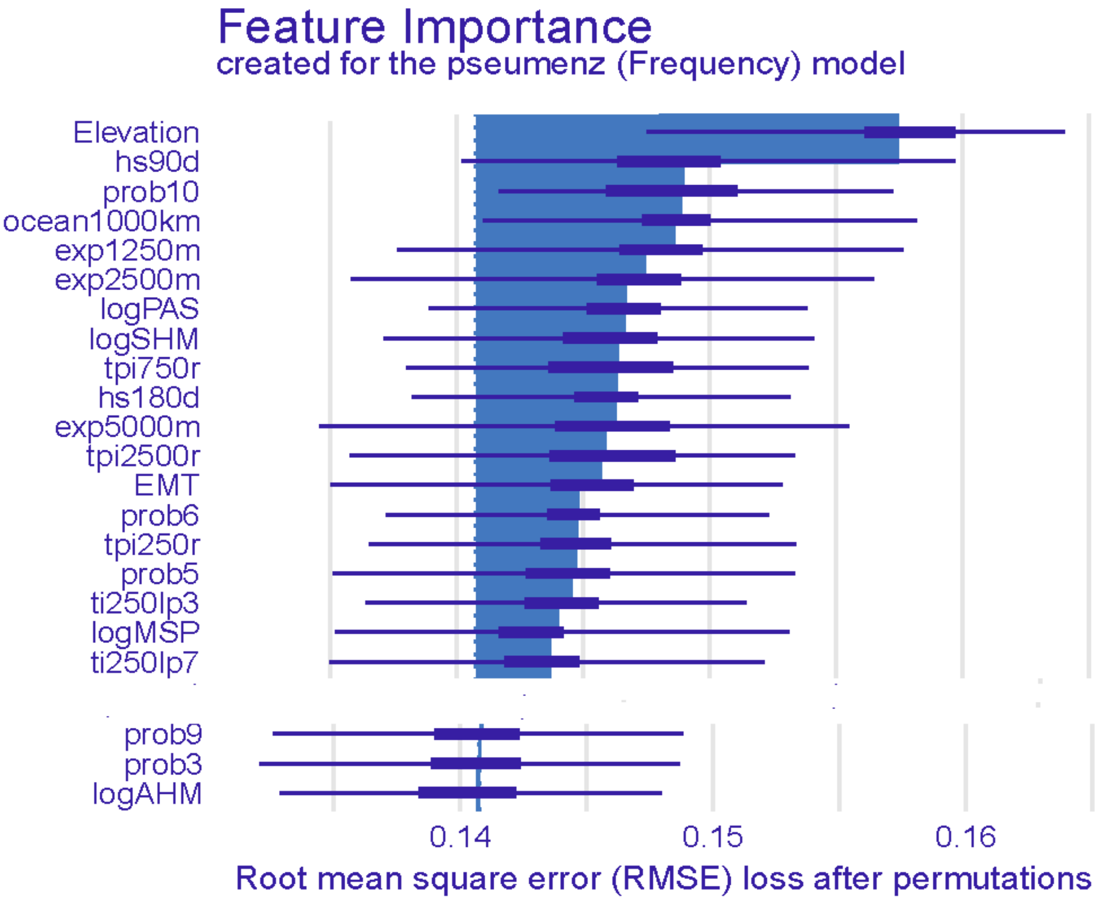

Fig. 12 Variable importance of predictor variables (y-axis) by root mean square error loss after removal of variable for frequency model (DNN 2a) of Douglas fir.

.

3.3. Model performance metrics and effects of predictor inclusion

Quantitative performance metrics confirm that the inclusion of additional predictor groups improved model performance across species. Mean absolute error (MAE) values calculated on withheld data decreased consistently as topographic variables and land cover probabilities were added to climate-only models (Table 1). Across the four species evaluated, inclusion of topographic predictors reduced MAE by 21–31%, with further error reductions of 25–37% when probabilistic land cover variables were included.

Table 1 Mean Absolute Error scores (percent improvement from climate only) of four modelled species as variables were added to model training.

Species	**Climate Only**	**Climate & Topographic Variables**	**Added Land Cover Probabilities**
Subalpine fir	0.93	0.67 (28%)	0.59 (37%)
Black spruce	0.75	0.59 (21%)	0.56 (25%)
Trembling aspen	0.99	0.75 (24%)	0.70 (29%)
Douglas fir	1.61	1.11 (31%)	1.03 (35%)

Absolute MAE values remain relatively high due to the zero-inflated nature of the response variable, as most plots contain no observations of a given species. As a result, errors reflect both incorrect predictions of presence and errors in predicted frequency where species are present. The primary purpose of these metrics is therefore comparative rather than absolute: to demonstrate that successive additions of ecologically meaningful predictors systematically improved model performance. These improvements are consistent with the variable importance results (Figs. 11–12) and the spatial patterns observed in the species distribution maps (Figs. 9–10).

3.4. Training behaviour and model stability

Model training dynamics further indicate stable learning behaviour. Across species, loss values declined smoothly over training epochs and plateaued within approximately 15 epochs (Fig. 13). Early stopping at this point prevented overfitting while retaining predictive performance. The absence of erratic loss behaviour or divergence suggests that the preprocessing, scaling, and data preparation steps described earlier were effective in supporting stable neural network training at continental scales.

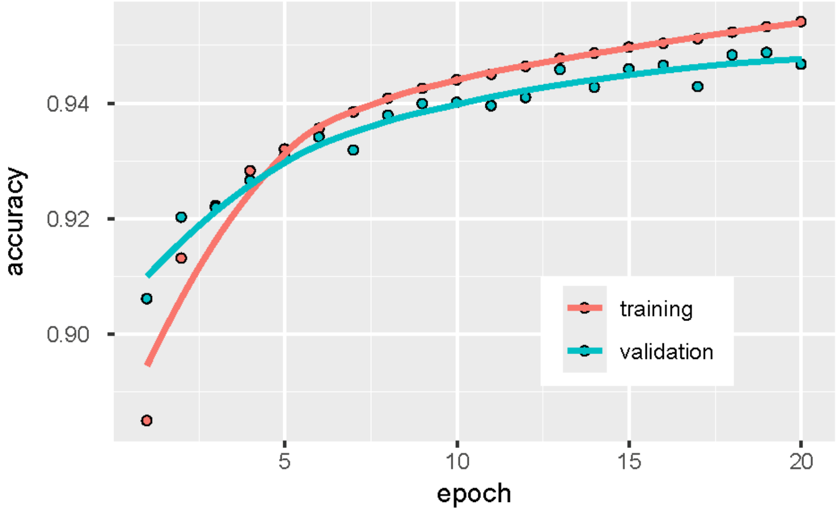

Fig. 13 DNN model performance over time

3.5. Limitations and future improvements

Several aspects of the framework warrant further development to strengthen its applicability and generality. First, although the modelling approach is designed to scale to a large number of species, detailed evaluation in this study focused on a limited set of representative taxa. Extending systematic evaluation to a broader range of species spanning different functional types, range sizes, and inventory representation would provide a more comprehensive assessment of model behaviour and robustness.

Second, model validation relied primarily on internal random data partitioning and qualitative comparison with known biogeographic patterns. Additional validation strategies, such as spatially stratified cross-validation across ecoregions or climatic gradients, would provide stronger tests of spatial transferability and help quantify uncertainty in extrapolation beyond well-sampled regions.

Third, future development of the framework could incorporate soil variables to better represent edaphic constraints on species distributions, using recently available gridded soil datasets at appropriate spatial scales, such as the 250 m SoilGrids products (Hengl et al. 2017; Poggio et al. 2021). Because climate, topography, land cover, and soil predictor layers are often derived from overlapping environmental covariates and spatial information, partial autocorrelation among predictor groups is expected. Post hoc analyses that partition shared and unique contributions to model accuracy, such as variance partitioning or commonality analysis, could therefore help clarify the relative influence of different predictor groups and improve interpretability of model outcomes.

Together, these future directions reflect opportunities to extend and strengthen the framework presented here. The results of this study already demonstrate that integrating harmonized forest inventory data with climate, topographic, and probabilistic land cover predictors in a deep learning framework is a viable and promising basis for large-scale species frequency modelling.

4. References

Alberta Environment & Parks. 2021. Ecological Site Information System (ESIS). Available from https://www.alberta.ca/esis.aspx.

Beaudoin, A., Bernier, P.Y., Guindon, L., Villemaire, P., Guo, X.J., Stinson, G., Bergeron, T., Magnussen, S., and Hall, R.J. 2014. Mapping attributes of Canada's forests at moderate resolution through

Booth, T.H. 2018. Species distribution modelling tools and databases to assist managing forests under climate change. Forest Ecology and Management **430**: 196-203.

Botella, C., Joly, A., Bonnet, P., Monestiez, P., and Munoz, F. 2018. A Deep Learning Approach to Species Distribution Modelling. Multimed Syst Appl: 169-199.

Chapin, F.S., Randerson, J.T., McGuire, A.D., Foley, J.A., and Field, C.B. 2008. Changing feedbacks in the climate-biosphere system. Frontiers in Ecology and the Environment **6**(6): 313-320.

Comisión Nacional Forestal. 2020. Inventario Nacional Forestal y de Suelos (INFS). Available from https://www.gob.mx/conafor.

Dar, S.A., Nabi, M., Dar, S.A., and Ahmad, W.S. 2022. Influence of anthropogenic activities on the diversity of forest ecosystems. *In* Towards sustainable natural resources: Monitoring and managing ecosystem biodiversity. Springer International Publishing, Cham. pp. 33-49.

Esquivel-Muelbert, A., Baker, T.R., Dexter, K.G., et.al. 2019. Compositional response of Amazon forests to climate change. Global Change Biology **25**(1): 39-56.

Evans, J.S., Murphy, M.A., Holden, Z.A., and Cushman, S.A. 2011. Modeling Species Distribution and Change Using Random Forest. Predictive Species and Habitat Modeling in Landscape Ecoloogy: Concepts and Applications: 139-159.

Hengl, T., de Jesus, J.M., Heuvelink, G.B.M., Gonzalez, M.R., Kilibarda, M., Blagotic, A., Shangguan, W., Wright, M.N., Geng, X.Y., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B. 2017. SoilGrids250m: Global gridded soil information based on machine learning. Plos One **12**(2).

LeCun, Y., Bengio, Y., and Hinton, G. 2015. Deep learning. Nature **521**(7553): 436-444.

Lee-Yaw, J.A., McCune, J.L., Pironon, S., and Sheth, S.N. 2022. Species distribution models rarely predict the biology of real populations. Ecography **2022**(6).

Little, E.L. 1971. Atlas of United States Trees. U.S. Department of Agriculture, Forest Service, Washington, D.C.

Martin, T.G., Wintle, B.A., Rhodes, J.R., Kuhnert, P.M., Field, S.A., Low-Choy, S.J., Tyre, A.J., and Possingham, H.P. 2005. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecol Lett **8**(11): 1235-1246.

Massey, R., Rogers, B.M., Berner, L.T., Cooperdock, S., Mack, M.C., Walker, X.J., and Goetz, S.J. 2023. Forest composition change and biophysical climate feedbacks across boreal North America. Nature Climate Change **13**(12): 1368-1375.

Meidinger, D., and Pojar, J. 1991. Ecosystems of British Columbia. BC Ministry of Forests, Research Branch.

National Forest Inventory. 2024. Multi-Agency Ground Plot (MAGPlot) Database – Standards for Forest Ground Plot Data Harmonization, Version 1.0. Available from https://open.canada.ca/data/en/dataset/8824392d-464e-413d-8bde-eaed61c79743.

NRCAN. 2021. Canada’s National Forest Inventory – Ground Plot Compilation Standards, Version 2.4. Available from http://nfi.nfis.org.

Ohmann, J.L., Gregory, M.J., Henderson, E.B., and Roberts, H.M. 2011. Mapping gradients of community composition with nearest-neighbour imputation: extending plot data for landscape analysis. Journal of Vegetation Science **22**(4): 660-676.

Pan, Y.D., Birdsey, R.A., Fang, J.Y., Houghton, R., Kauppi, P.E., Kurz, W.A., Phillips, O.L., Shvidenko, A., Lewis, S.L., Canadell, J.G., Ciais, P., Jackson, R.B., Pacala, S.W., McGuire, A.D., Piao, S.L., Rautiainen, A., Sitch, S., and Hayes, D. 2011. A Large and Persistent Carbon Sink in the World's Forests. Science **333**(6045): 988-993.

Poggio, L., de Sousa, L.M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., and Rossiter, D. 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil-Germany **7**(1): 217-240.

Rozenberg, G.S. 2010. Mixed Effects Models and Extensions in Ecology with. Uchen Zap Kaz U-Este **152**(3): 275-278.

U. S. Forest Service. 2023. Forest Inventory and Analysis National Program. Available from https://www.fia.fs.usda.gov.

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

Wilson, B.T., Lister, A.J., and Riemann, R.I. 2012. A nearest-neighbor imputation approach to mapping tree species over large areas using forest inventory plots and moderate resolution raster data. Forest Ecol Manag **271**: 182-198.