Title

Sarah J. Hart, Asha Paudel, Josh Carrell, Maxwell Cook

Sept 20, 2023

# Abstract

# Highlights

* Across the Southern Rocky Mountains, aspen occupies a wide climatic niche.

# Introduction

Specifically our objectives are to: (1) better understand the relationships between aspen presence and climate, topographic, and edaphic factors and (2) map the area suitable for aspen under current climate conditions, and (3) project areas where aspen may expand, contract, or remain stable.

# Materials and Methods

## Study area

The study area consists of the Southern Rocky Mountains Ecoregion (SRME), an area of approximately 145,700 km2 that extends from southern Wyoming to northern New Mexico (Fig. 1. The SRME consists of rugged, topography with elevation ranging from 1450 m to above 4400 m, seven mountain ranges that largely trend north-south, and four Intermontane basins ([Drummond, 2012](#ref-drummond2012SouthernRockiesEcoregion)). The climate of the SRME is characterized by a continental climate, with hot summers (mean maximum July temperature of XX\*C) and cool winters (mean minimum January temperatures of XX\*C), and intermediate precipitation (mean annual precipitation of XXX mm), most of which falls as snow [Lukas et al. ([2014](#ref-lukas2014ClimateChangeColorado)); @rodman2021EffectsBarkBeetle]. At local scales, the climate is driven by elevation gradients, the prevailing westerly winds, and the north-south orientation of the mountain. Thus temperature and precipitation can vary dramatically over short distances. Temperature is generally greatest at low elevations, while precipitation is generally greatest at higher elevations, particularly on the windward side of the Rockies ([Lukas et al., 2014](#ref-lukas2014ClimateChangeColorado)). Precipitation patterns exhibit a distinct latitudinal gradient, where more southern locations often experience more precipitation in the summer months due to the North American Monsoon system ([Lukas et al., 2014](#ref-lukas2014ClimateChangeColorado)).

Ecosystems of the SRME correspond with topoclimatic patterns; warm and dry low elevation valleys and intermountain basins are dominated by grasslands and shrublands, forest occupy intermediate elevations, while grasses, sedges, cushion plants, forbs, mosses, and lichens dominate cold alpine elevations. Within the ca. 55% of the SRME that is forested (REF), topoclimatic variation drives tree species composition. Lower Montane forests (< 2,300 m) are generally composed of ponderosa pine (*Pinus ponderosae*) woodlands, piñon (*Pinus edulis*) and juniper (*Juniperus* spp.) woodlands, and gambel oak (*Quercus gambelii*) shrublands. Forests of the Upper Montane zone (ca. 2,300 - 2,800 m) are dominated by ponderosa pine-Douglas fir mixed conifer systems, quaking aspen, and lodgepole pine (*Pinus contorta*). Forests of the subalpine zone (ca. 2,800 m - 3,200 m) are dominated by Engelmann spruce, subalpine fir, and to a lesser limber pine (*Pinus flexilis)* and Rocky Mountain bristlecone pine (*P. aristata).* Forests dynamics across the SRME are strongly shaped by climate-sensitive disturbances, notably wildfires, outbreaks of native bark beetles, and windstorms ([Baker and Veblen, 1990](#ref-baker1990); [Peet, 1981](#ref-peet1981ForestVegetationColorado); [Veblen et al., 2000](#ref-veblen2000ClimaticHumanInfluences); [Veblen et al., 1994](#X6d15166298ccfaa854e56f5197e8b0dbab0811c)).

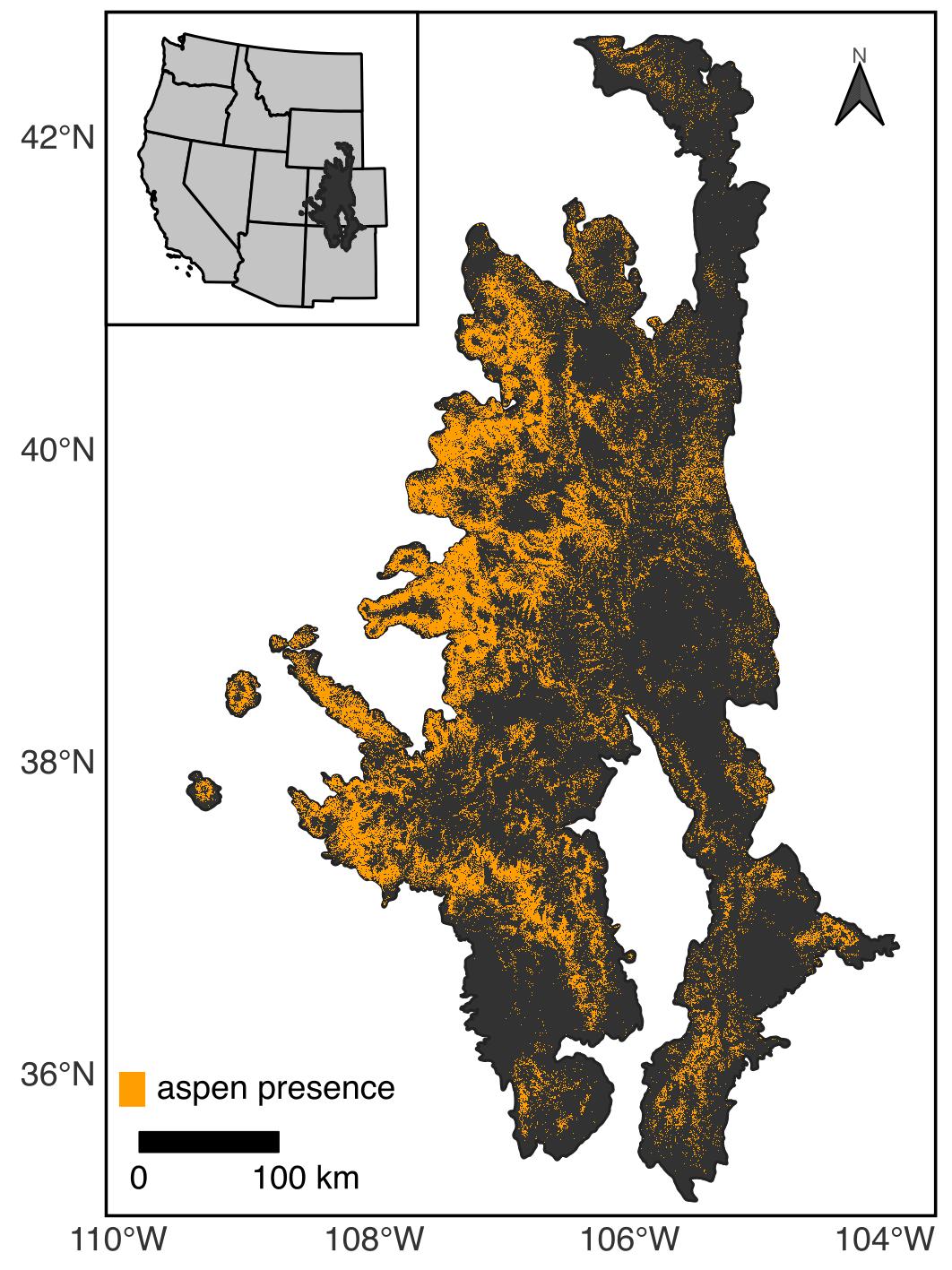


Figure 1: The Southern Rocky Mountain Ecoregion and current distribution of aspen.

## Data

### Species occurrence data

### Predictor variables

To understand and project the relationship between aspen and climate under historical and future conditions, we obtained gridded climate data from the AdaptWest Project ([2022](#Xcb40ace7ff505c24620413cbf7691d7f26a4667)), which provides both current and future climate data that are downscaled to 1 x 1km resolution using the ClimateNA software (version 7.3) ([Wang et al., 2016](#ref-wang2016LocallyDownscaledSpatially)) . Current conditions, defined here as climatalogical norms for the 1981-2010 period, were generated from 4 x 4 km climate data provided by the PRISM Climate Group ([2021](#ref-prismclimategroup2021)). Future climate conditions were generated from data included in the sixth phase of Coupled Model Intercomparison Project (CMIP6). Here we used projections of future climate for the periods 2011-2040, 2041-2070, and 2071-2100. Given considerable uncertainty about future emissions, we used data depicting two scenarios (i.e., Shared Socioeconomic Pathways; SSPs) generated under CMIP6, SSP2-4.5 and SSP5-8.5. The SSP2-4.5 scenario describes an intermediate scenario characterized by moderate increases in emission through 2040 followed by a decline, while the SSP5-8.5 scenario describes a more extreme situation where emissions increase through 2100 ([Riahi et al., 2017](#ref-riahi2017SharedSocioeconomicPathways)). In addition to uncertainty about societal decisions about greenhouse gas emissions represented in the SSPs, considerable variation exists among the more than 50 atmosphere-ocean general circulation models (AOGCMs) included in CMIP6 due differences in complexity, assumptions, and parameterization. At the global scale, predictions made by some AOGCMs are believed to overestimate future warming ([Hausfather et al., 2022](#X0773c8e2ab11f078ec717d59314f0220fe06795)). While at more regional scales, AOCGMs differ in the number of simulations for the historical period, amonunt bias, and spatial grain [REF]. Here we make use of an ensemble data constructed from eight AOCGMs identified by (REF) as being appropriate for regional applications, such as species distribution modeling, across North America (REF).

*something about how lots of different climate variables exist* and best practices.

The AdaptWest Project ([2022](#Xcb40ace7ff505c24620413cbf7691d7f26a4667)) provides data for 33 different biologically relevant climate variables (Table ??. In addition to these variables, we calculated five other variables that have been identified as important predictors of aspen habitat ([Rehfeldt et al., 2009](#ref-rehfeldt2009AspenClimateSuddena); [Rehfeldt et al., 2006](#X5be6899825a8c3d2d140aa48216acf46924864d)), an annual dryness index (ADI), the total growing season (April - August) precipitation (GSP), the ratio of GSP to annual precipitation (PRATIO), the mean maximum temperature in warmest month (TMAX), and a growing season precipitation to degree day ratio (GSPDD5) (Table ??.

Given mountainous areas such as the Southern Rocky Mountains are characterized high topoclimatic variation ([Franklin et al., 2013](#ref-franklin2013ModelingPlantSpecies)), we further downscaled all climate variables from a 1 km resolution to a 250 m resolution using gradient and inverse distance squared (GIDS) interpolation ([Flint and Flint, 2012](#ref-flint2012DownscalingFutureClimate); [Nalder and Wein, 1998](#X44a33479d922de88fa614470bb4d0b1d1f5aa3e)), following methods outlined in Rodman et al. ([2020](#ref-rodman2020ChangingClimateSnuffing)). As ancillary data in the downscaling, we used a 30-m digital elevation model (DEM) from the USGS (ref).

In addition to climate variables, we also included data describing terrain and soils as predictors in our models. To account for the potential effects of local topographic variation on soil transport and water balance, we use the DEM to calculate the topographic position index [TPI; Weiss ([2001](#ref-weiss2001))]. We calculated TPI for a 3-cell neighborhood (TPI3) and a 15-cell neighborhood (TPI15) to represent both fire and coarse scale topographic patterns ([Rodman et al., 2020](#ref-rodman2020ChangingClimateSnuffing)). To account for the effects of aspect and slope on microclimate, we calculated the Heat Load Index [HLI; McCune and Keon ([2002](#ref-mccune2002EquationsPotentialAnnual)); McCune ([2007](#ref-mccune2007ImprovedEstimatesIncident))]. Both HLI and TPI were calculated in R using the spatialeco package([Evans and Murphy, 2021](#ref-spatialEco)). We also included soil properties in our models of aspen habitat suitability. Specifically, we obtained 30-m probabilistic maps of soil pH, the percentage of organic material, the percentage of clay, and saturated soil water content from the POLARIS database ([Chaney et al., 2019](#ref-chaney2019POLARISSoilProperties)).

We did not include elevation, latitude, and longitude in our modeling because relationships are expected to be only correlative and thus may affect predictor explanation and projection ([Araújo et al., 2019](#Xa394076d2f0f03ea6b9000cd6143e30f891c94f)). Instead we compare model residuals with these variables.

Table 1: Predictor variables selected for modeling the distribution of aspen

| Variable | Description | Expected relationship with aspen |
| --- | --- | --- |
| Climate variables |  |  |
| ADI | annual dryness index: (degree-days above 5 °C)^0.5 / (mean annual precipitation) | Higher ADI is expected to limit aspen (Rehfeltd et al. 2009) |
| DD\_0 | degree-days below 0 °C | Very cold temperatures may limit aspen (Rehfeltd et al. 2009) |
| MAR | mean annual solar radiation (MJ m‐2 d‐1) | Greater MAR may inhibit aspen, particularly at warmer sites |
| PRATIO | mean precipitation ratio: (growing season precipitation) / (mean annual precipitation) | Evenly distributed precipitation (intermediate PRATIO) may promote aspen by limiting seasonal moisture stress (Rehfeltd et al. 2009) |
| RH | mean annual relative humidity (%) | Higher relative humidity may limit seasonal moisture stress. |
| Soil variables |  |  |
| OM | soil organic matter [log10(%)] | Aspen is expected grow better on soils with high organic matter content (Perala 1990) |
| SWC | saturated water content (m3/m3) | Aspen is expected grow better on soils with greater water holding capacity (Perala 1990) |
| Clay | clay | High clay contents may inhibit aspen growth (Jones and DeByle 1985) |
| Topographic variables |  |  |
| HLI | heat load index | Greater HLI may inhibit aspen, particularly at warmer sites |
| TPI3 | topographic position index calculated using a 3 x 3 cell neighborhood | Aspen is expected grow better in valley bottoms (low TPI) and on benches (moderate TPI) than steep slopes (high TPI) (Jones and DeByle 1985) |

### Data processing

## Modeling Approach

#### Overview

To characterize suitable habitat for aspen, we used three different modeling approaches commonly applied in species distribution modeling, generalized linear models (GLMs), gradient boosted tree (GBTs), and random forests (RFs). We additionally fit generalized additive models (GAMS), however preliminary analyses revealed both a poor fit and response-predictor relationships that did not align with hypothesized relationships (Appendix 10. Here, we first overview our modeling approach before describing specifics for individual modeling techniques. All models were fit in *R* ([R Core Team, 2022](#X4878592beea4a6dfca9c91383c925df652b2c7a)) using the *tidymodels* framework ([Kuhn and Wickham, 2020](#ref-tidymodels)).

To build GLM, RF, and GBT models, we first constructed a balanced data consisting of 10,000 pixels with aspen present and 10,000 pixels without aspen. To minimize the potential effects of spatial autocorrelation, pixels were selected so that they were separated by at least 1 km. To improve modeling fit and interpretability, all predictor variables were standardized by calculating the z-score. Using this dataset, we reduced our set of topoclimatic predictors to minimize the potential effects of collinearity on model inference and projection. Specifically, we used the *spatialRF* ([Wright and Ziegler, 2017](#ref-spatialRF)) to identify variables with a variable inflation factor greater than 5 or a correlation of greater than 0.70 with any other variable. When these conditions were met, variables were removed preferentially (Table 1.

To build and evaluate models, we then split our dataset into equally-sized testing and training datasets. The testing dataset was further split to create five spatial cross-validation folds using the *spatialsample* package ([Mahoney et al., 2023](#ref-spatialsample)). We then tuned model hyperparameters using spatial cross-validation and identified the best parameters based on the area under the receiver operator curve. To evaluate model performance in three ways, we used three approaches, but for each approah calculated the following statistics: XX.. First, we used spatial cross-validation to evaluate the capacity of our model to predict to new areas. Second, we evaluated performance on the testing dataset. Third, we compared the predictions with an independent data set of aspen presence. In all cases, we calculated

For each model we determined variable importance using a model-agnostic permutation-based approach, where each variable is randomized and then the ROC AUC statistic is compared with ROC AUC for the full model (where data has not been randomized). We evaluated the relationship between aspen presence and each predictor variable using accumulated local effects (ALE) profiles. Both variable importance and ALE were calculated in R using the DALEX package ([Biecek, 2018](#ref-DALEX)).

#### Generalized linear models

GLMs included a logit link function and a binomial error distribution and were fit using Lasso regularization approach available within the *glmnet* package. For all variables, we included both linear and quadratic effects. We did not explore any interaction terms.

#### Random Forests

RF models are an extension of classification and regression trees (CART). RFs are useful for modeling nonlinear relationships and complex interactions, which often characterize ecological relationships ([Cutler et al., 2007](#X8af443e7151d3081ddf81ae6488f6e02a99b7ea)). Several hyperparamaters can be tuned in RF models. Here we usedFor the RF model, we tuned the minimum number of data points in a node that is required for the node to be split further (min\_n) and the number of variables to try at each split (mtry). RF models were fit using the *ranger* package ([**ranger?**](#ref-ranger)).

#### Gradient boosting trees

GBTs are also For the GBT model, we tuned the minimum number of data points in a node that is required for the node to be split further (min\_n), the reduction in the loss function required to split further (loss\_reduction), the learning rate (learn\_rate) and the maximum depth of the tree (tree\_depth). Based on highest AUC, min\_n = XXX, loss\_reduction = XXX , learn\_rate = XXX , and tree\_depth = XXX. GBTs were fit using the package xgboost ([Chen et al., 2023](#ref-xgboost)).

#### Ensemble Model

We calculated a weighted probability of occurrence from all three presence-absence models. Weights assigned were based on the ROC AUC statistic.

## Projection

# Results

## Model performance

The ensemble model correctly predicted the presence of aspen for 83% of points identification points (n=14417.

(Table 2).

(Table 5).

Table 2: Model performance statistics. Observed values are from independent testing data and predicted value assume a modeled probabilty of aspen occurrence of at least 0.5

| Model | Accuracy | F measure | kappa | Precision | Recall | AUC ROC | Sensitivity | Specificity |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Ensemble | 0.78 | 0.78 | 0.56 | 0.78 | 0.79 | 0.87 | 0.79 | 0.78 |
| GLM | 0.75 | 0.74 | 0.51 | 0.78 | 0.71 | 0.83 | 0.71 | 0.80 |
| RF | 0.80 | 0.79 | 0.59 | 0.80 | 0.78 | 0.88 | 0.78 | 0.81 |
| XGB | 0.75 | 0.76 | 0.50 | 0.73 | 0.80 | 0.84 | 0.80 | 0.70 |

### Variable importance

(Fig. 2.

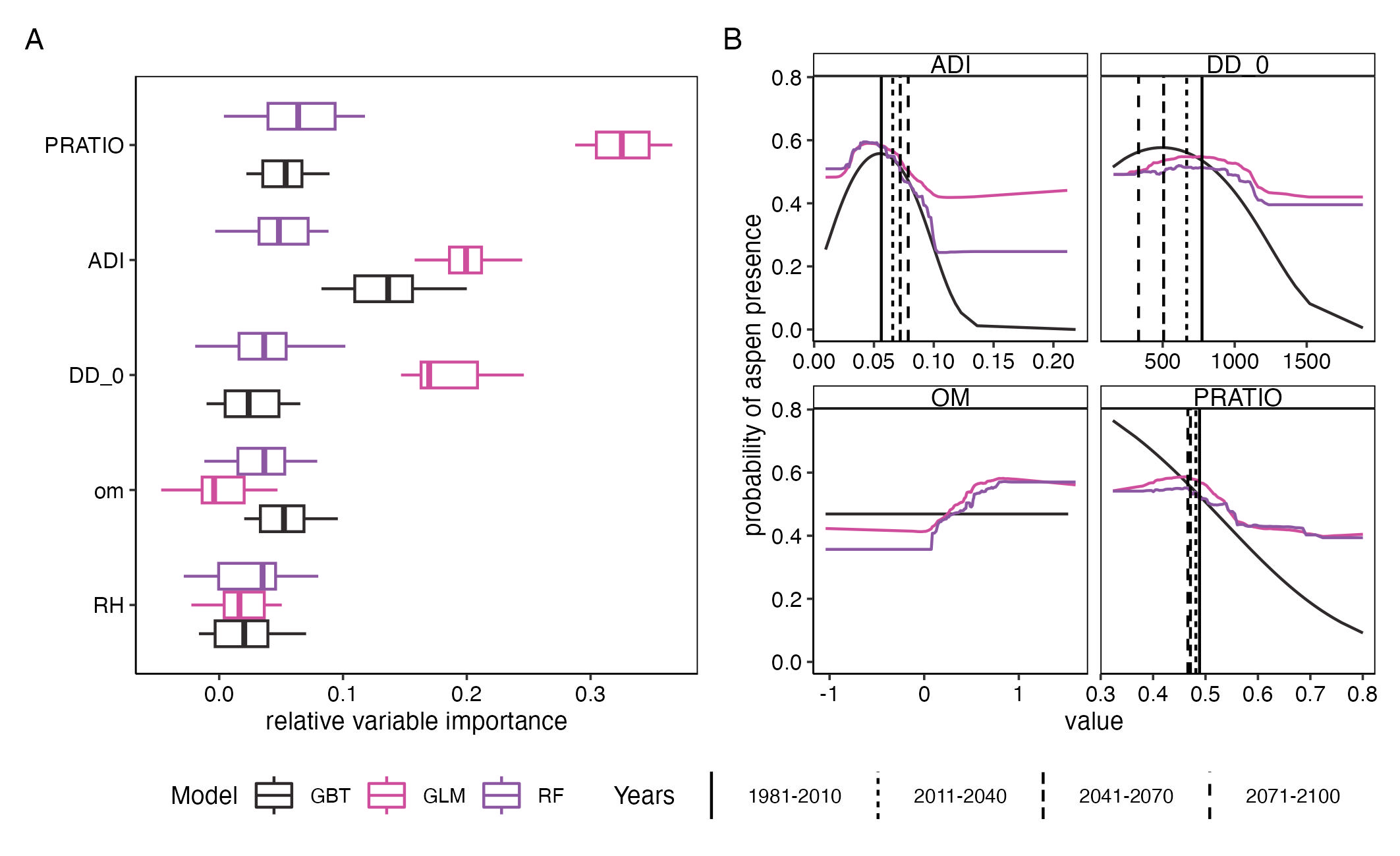


Figure 2: Variable importance

(Fig. 3.

(Fig. 4.

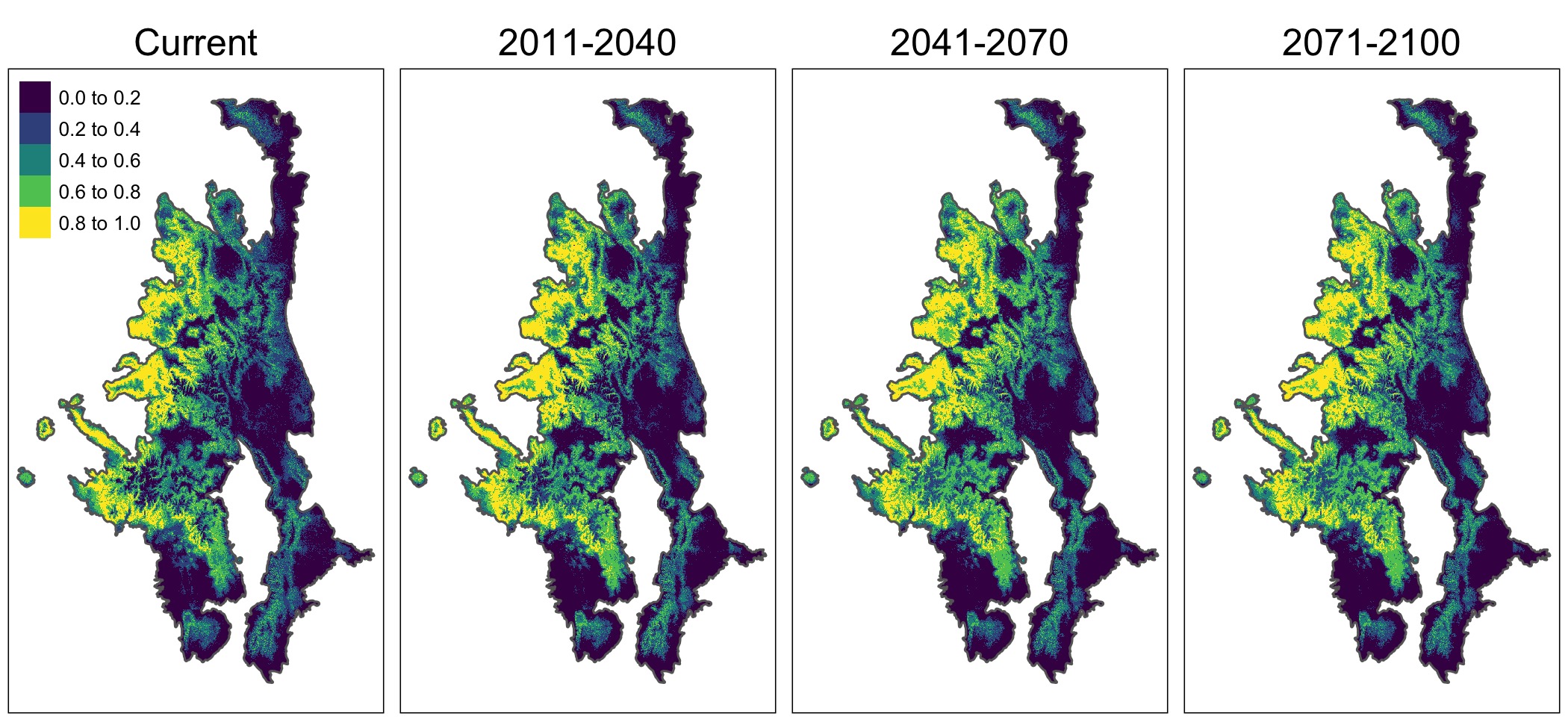


Figure 3: The ensemble projection of aspen habitat suitability under current conditions and projections for future periods based on an SSP5-8.5 scenario.

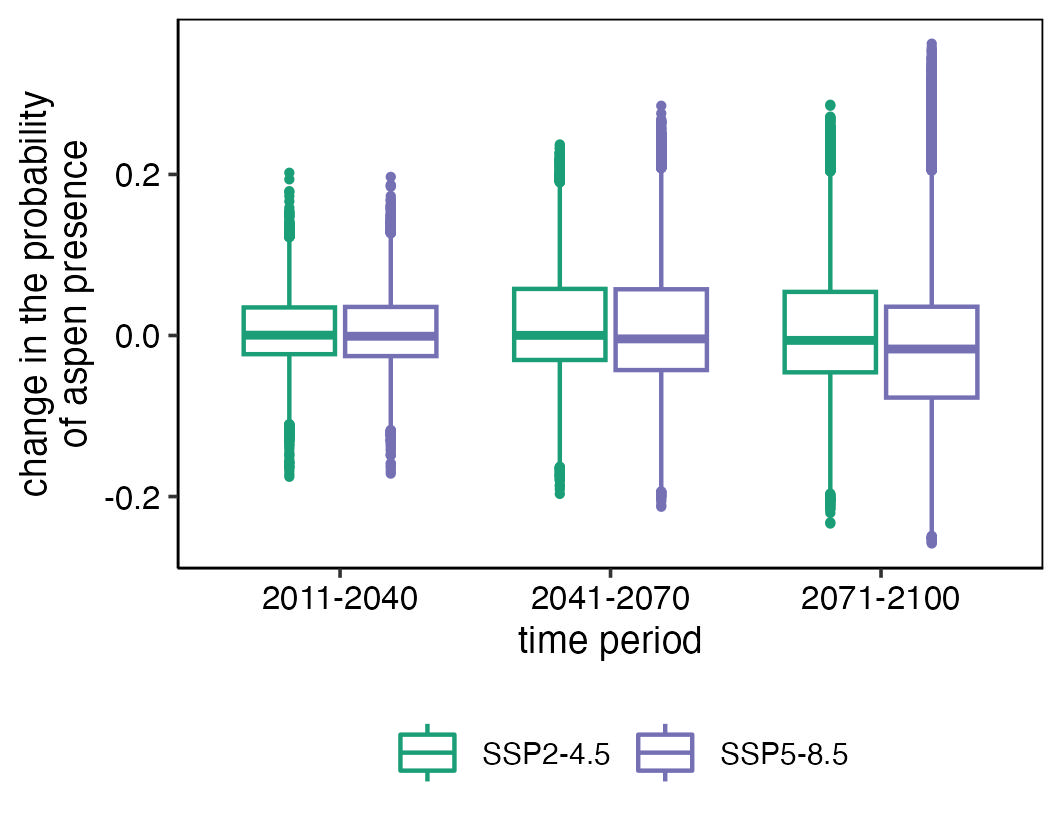


Figure 4: Change

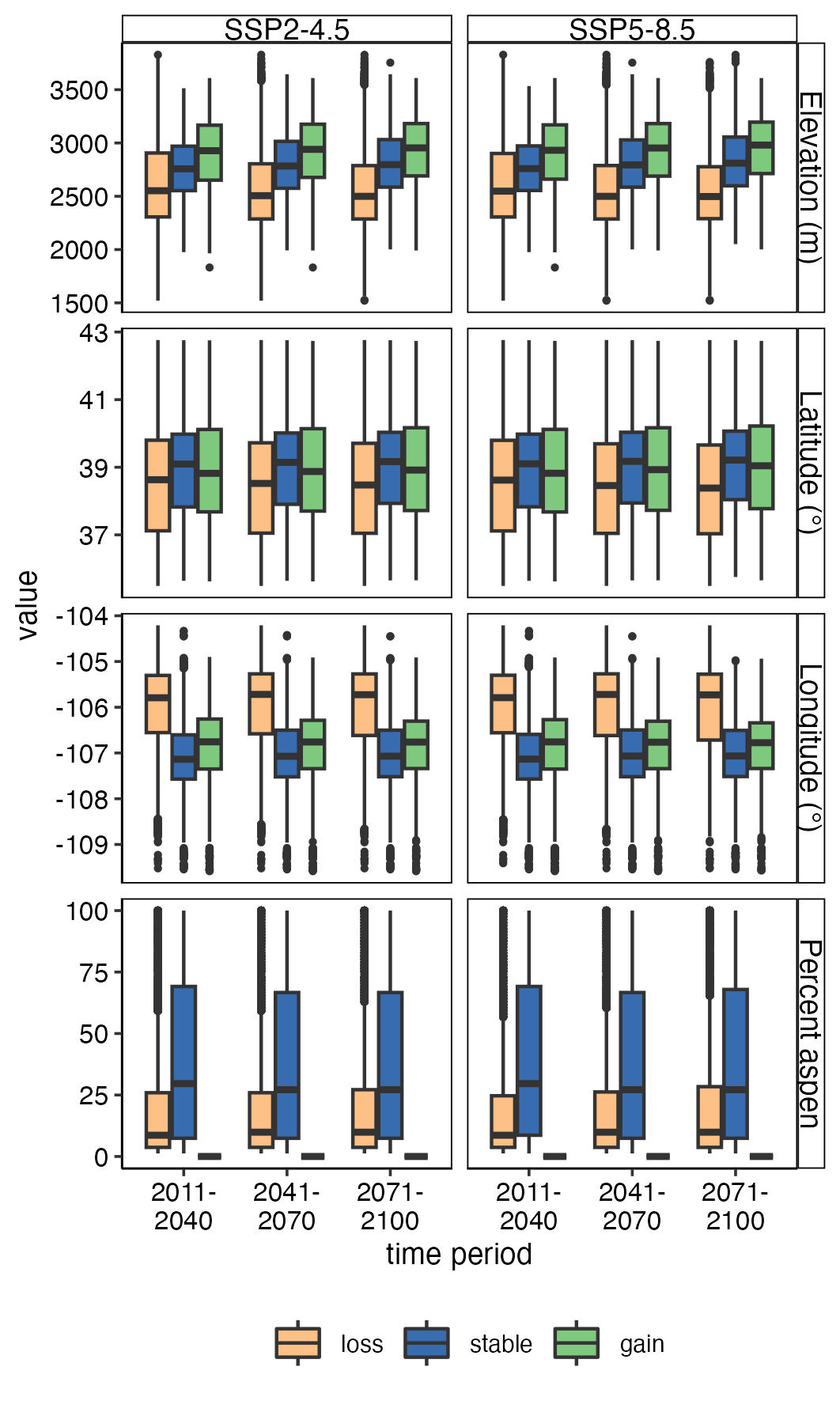


Figure 5: GainLoss-Geo

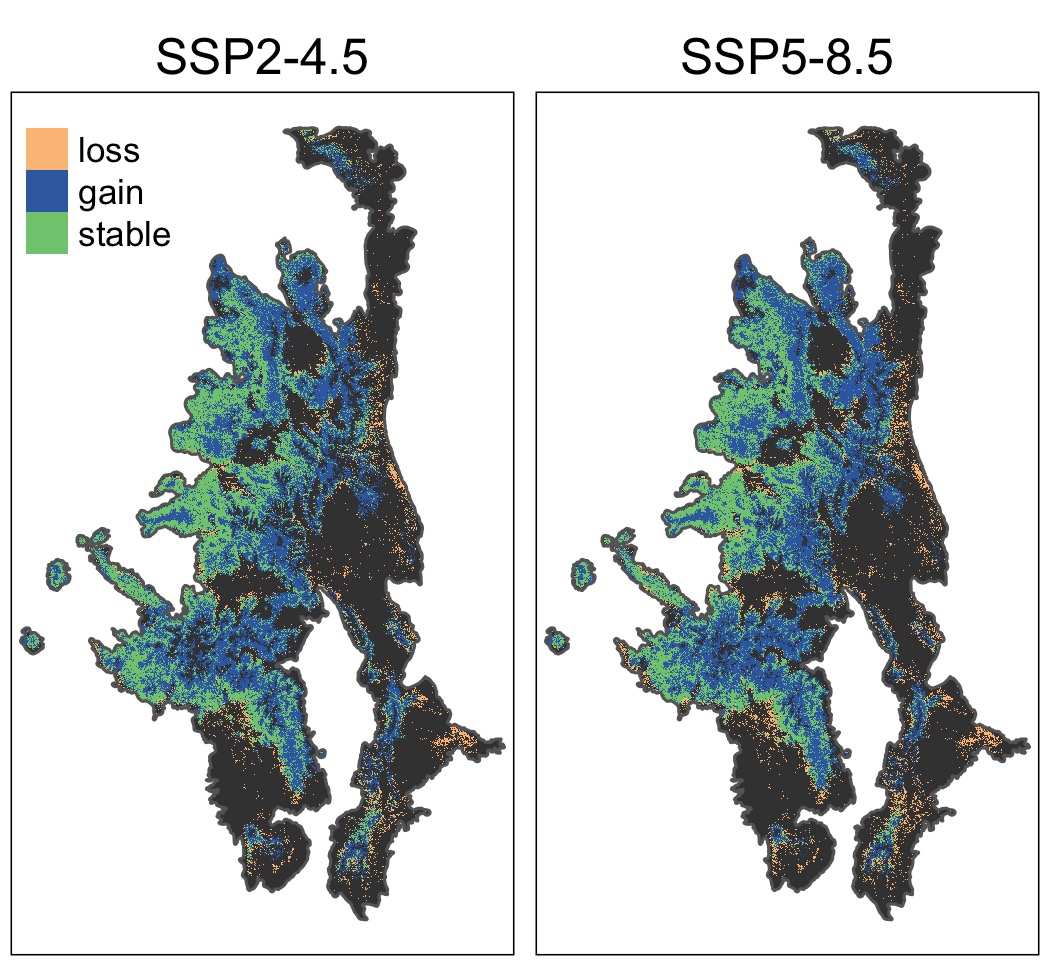


Figure 6: LossGainMaps

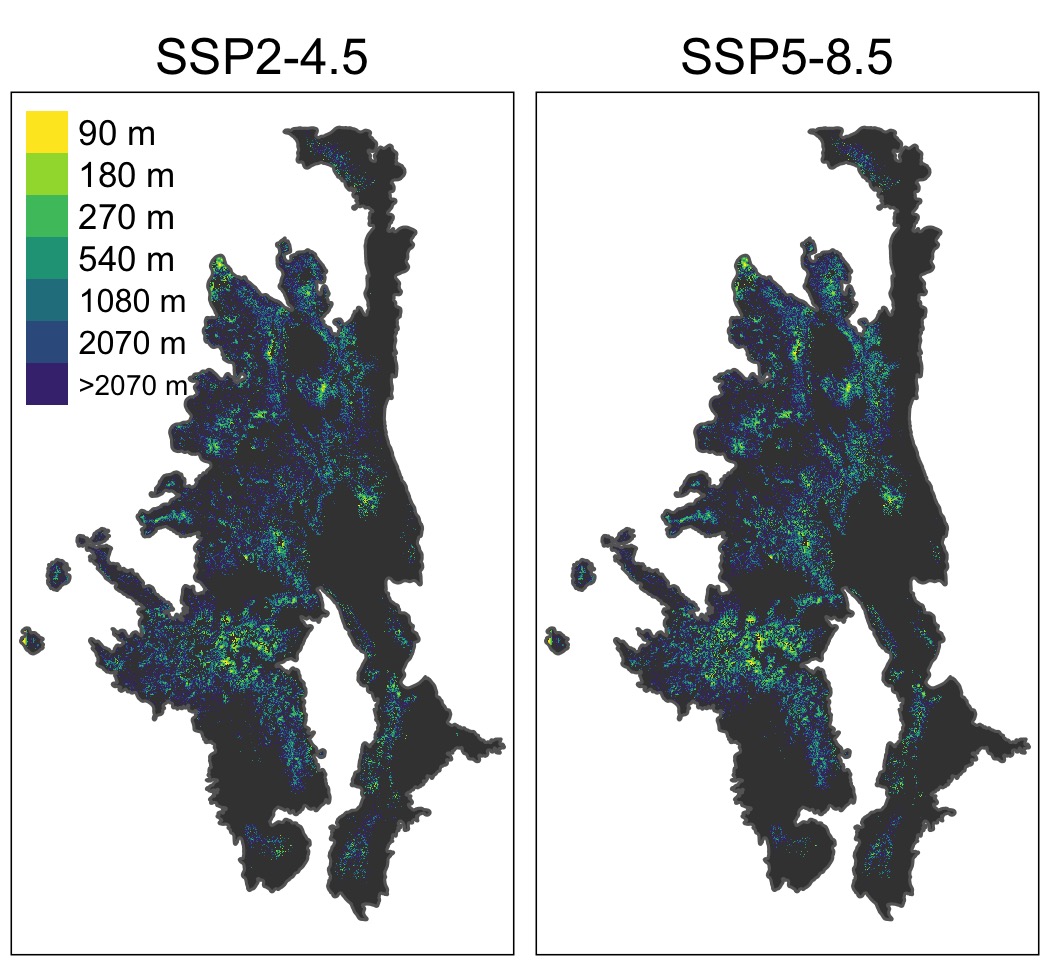


Figure 7: GainDistanceMaps

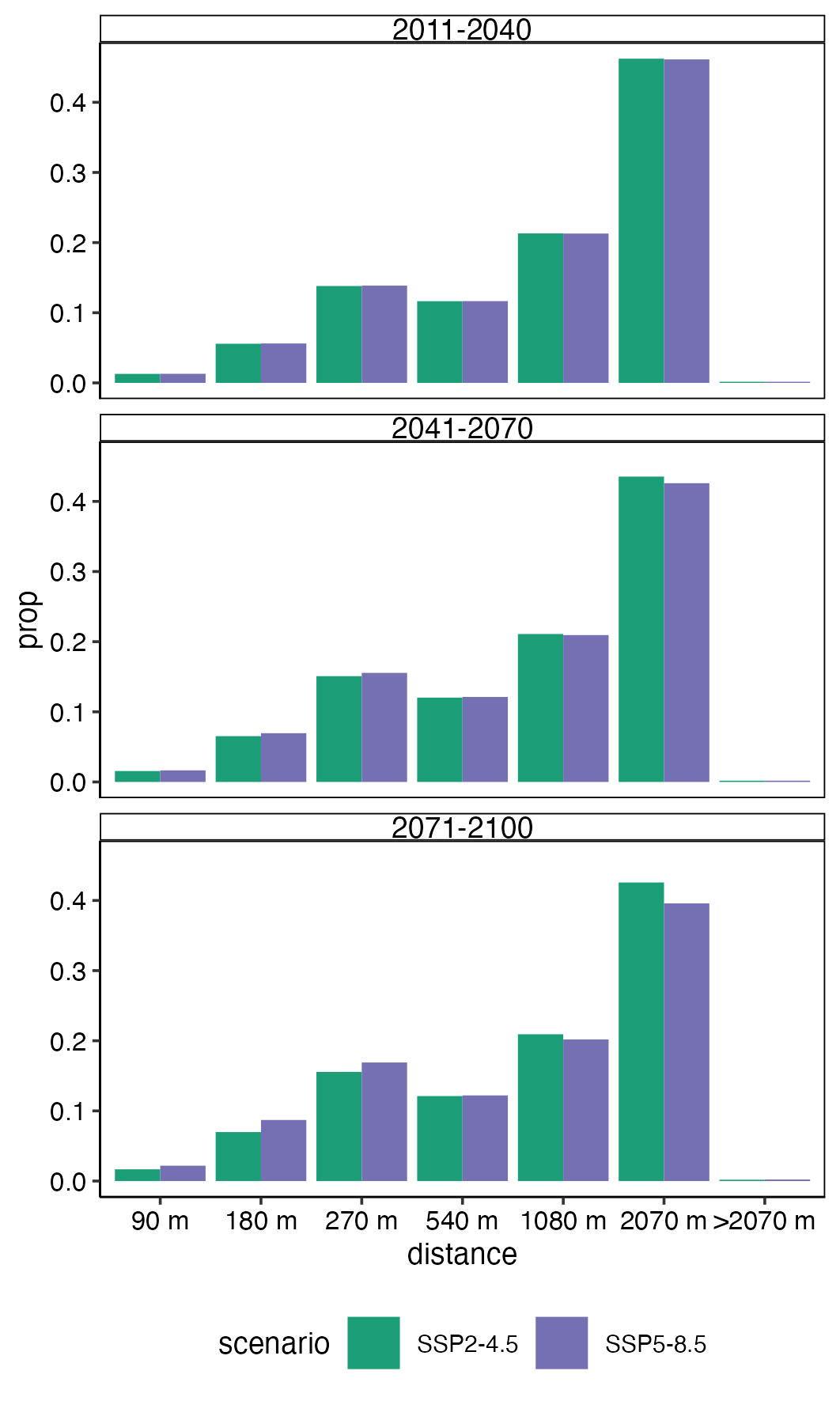


Figure 8: GainDistanceMaps

# Discussion

Quaking aspen is one of the most widely distributed tree species. While our study area clearly includes areas where aspen is absent, we acknowledge that our dataset may be environmentally truncated ([Hannemann et al., 2016](#ref-hannemann2016DevilDetailUnstable); [Thuiller et al., 2004](#Xb5e3a842768cfe7ba494facd83499a616e17792)).

# Conclusions

# References

AdaptWest Project, 2022. [Gridded current and projected climate data for north america at 1km resolution, generated using the ClimateNA v7.30 software (t. Wang et al., 2022).](https://Available at adaptwest.databasin.org.)

Araújo, M.B., Anderson, R.P., Márcia Barbosa, A., Beale, C.M., Dormann, C.F., Early, R., Garcia, R.A., Guisan, A., Maiorano, L., Naimi, B., O’Hara, R.B., Zimmermann, N.E., Rahbek, C., 2019. Standards for distribution models in biodiversity assessments. Science Advances 5, eaat4858. <https://doi.org/10.1126/sciadv.aat4858>

Baker, W.L., Veblen, T.T., 1990. Spruce beetles and fires in the nineteenth-century subalpine forests of western colorado, USA. Arctic and Alpine Research 22, 6580.

Biecek, P., 2018. [DALEX: Explainers for complex predictive models in r](https://jmlr.org/papers/v19/18-416.html) 19, 1–5.

Chaney, N.W., Minasny, B., Herman, J.D., Nauman, T.W., Brungard, C.W., Morgan, C.L.S., McBratney, A.B., Wood, E.F., Yimam, Y., 2019. POLARIS Soil Properties: 30-m Probabilistic Maps of Soil Properties Over the Contiguous United States. Water Resources Research 55, 2916–2938. <https://doi.org/10.1029/2018WR022797>

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., Yuan, J., 2023. [Xgboost: Extreme gradient boosting](https://CRAN.R-project.org/package=xgboost).

Cutler, D.R., Edwards Jr, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88, 27832792.

Drummond, M.A., 2012. [Southern rockies ecoregion: Chapter 8 in status and trends of land change in the western united states–1973 to 2000](http://pubs.er.usgs.gov/publication/pp1794A8). Reston, VA.

Evans, J.S., Murphy, M.A., 2021. [spatialEco](https://github.com/jeffreyevans/spatialEco).

Flint, L.E., Flint, A.L., 2012. Downscaling future climate scenarios to fine scales for hydrologic and ecological modeling and analysis. Ecological Processes 1, 2. <https://doi.org/10.1186/2192-1709-1-2>

Franklin, J., Davis, F.W., Ikegami, M., Syphard, A.D., Flint, L.E., Flint, A.L., Hannah, L., 2013. Modeling plant species distributions under future climates: how fine scale do climate projections need to be? Global Change Biology 19, 473–483. <https://doi.org/10.1111/gcb.12051>

Hannemann, H., Willis, K.J., Macias-Fauria, M., 2016. The devil is in the detail: unstable response functions in species distribution models challenge bulk ensemble modelling. Global Ecology and Biogeography 25, 26–35. <https://doi.org/10.1111/geb.12381>

Hausfather, Z., Marvel, K., Schmidt, G.A., Nielsen-Gammon, J.W., Zelinka, M., 2022. Climate simulations: recognize the ‘hot model’ problem. Nature 605, 26–29. <https://doi.org/10.1038/d41586-022-01192-2>

Kuhn, M., Wickham, H., 2020. [Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles.](https://www.tidymodels.org)

Lukas, J., Barsugli, J., Doesken, N., Rangwala, I., Wolter, K., 2014. Climate change in colorado: A synthesis to support water resources management and adaptation. University of Colorado, Boulder, Colorado.

Mahoney, M.J., Johnson, L.K., Silge, J., Frick, H., Kuhn, M., Beier, C.M., 2023. Assessing the performance of spatial cross-validation approaches for models of spatially structured data. <https://doi.org/10.48550/arXiv.2303.07334>

McCune, B., 2007. [Improved estimates of incident radiation and heat load using non-parametric regression against topographic variables](https://www.jstor.org/stable/4499284). Journal of Vegetation Science 18, 751–754.

McCune, B., Keon, D., 2002. [Equations for potential annual direct incident radiation and heat load](https://www.jstor.org/stable/3236745). Journal of Vegetation Science 13, 603–606.

Nalder, I.A., Wein, R.W., 1998. Spatial interpolation of climatic Normals: test of a new method in the Canadian boreal forest. Agricultural and Forest Meteorology 92, 211–225. <https://doi.org/10.1016/S0168-1923(98)00102-6>

Pedersen, E.J., Miller, D.L., Simpson, G.L., Ross, N., 2018. Hierarchical generalized additive models: an introduction with mgcv. <https://doi.org/10.7287/peerj.preprints.27320v1>

Peet, R.K., 1981. Forest vegetation of the colorado front range. Vegetatio 45, 375.

PRISM Climate Group, 2021. [Monthly 30-year climate normals (1981-2010)](https://prism.oregonstate.edu/normals/).

R Core Team, 2022. [R: A language and environment for statistical computing](http://www.R-project.org). R Foundation for Statistical Computing, Vienna, Austria.

Rehfeldt, G.E., Crookston, N.L., Warwell, M.V., Evans, J.S., 2006. Empirical analyses of plant-climate relationships for the western united states. International Journal of Plant Sciences 167, 11231150.

Rehfeldt, G.E., Ferguson, D.E., Crookston, N.L., 2009. Aspen, climate, and sudden decline in western USA. Forest Ecology and Management 258, 2353–2364. <https://doi.org/10.1016/j.foreco.2009.06.005>

Riahi, K., Vuuren, D.P. van, Kriegler, E., Edmonds, J., O’Neill, B.C., Fujimori, S., Bauer, N., Calvin, K., Dellink, R., Fricko, O., Lutz, W., Popp, A., Cuaresma, J.C., Kc, S., Leimbach, M., Jiang, L., Kram, T., Rao, S., Emmerling, J., Ebi, K., Hasegawa, T., Havlik, P., Humpenöder, F., Da Silva, L.A., Smith, S., Stehfest, E., Bosetti, V., Eom, J., Gernaat, D., Masui, T., Rogelj, J., Strefler, J., Drouet, L., Krey, V., Luderer, G., Harmsen, M., Takahashi, K., Baumstark, L., Doelman, J.C., Kainuma, M., Klimont, Z., Marangoni, G., Lotze-Campen, H., Obersteiner, M., Tabeau, A., Tavoni, M., 2017. The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: An overview. Global Environmental Change 42, 153–168. <https://doi.org/10.1016/j.gloenvcha.2016.05.009>

Rodman, K.C., Veblen, T.T., Battaglia, M.A., Chambers, M.E., Fornwalt, P.J., Holden, Z.A., Kolb, T.E., Ouzts, J.R., Rother, M.T., 2020. A changing climate is snuffing out post-fire recovery in montane forests. Global Ecology and Biogeography geb.13174. <https://doi.org/10.1111/geb.13174>

Thuiller, W., Brotons, L., Araújo, M.B., Lavorel, S., 2004. Effects of restricting environmental range of data to project current and future species distributions. Ecography 27, 165–172. <https://doi.org/10.1111/j.0906-7590.2004.03673.x>

Veblen, T.T., Hadley, K.S., Nel, E.M., Kitzberger, T., Reid, M., Villalba, R., 1994. Disturbance regime and disturbance interactions in a rocky mountain subalpine forest. Journal of Ecology 82, 125–135. <https://doi.org/10.2307/2261392>

Veblen, T.T., Kitzberger, T., Donnegan, J., 2000. Climatic and Human Influences on Fire Regimes in Ponderosa Pine Forests in the Colorado Front Range. Ecological Applications 10, 1178–1195. https://doi.org/<https://doi.org/10.1890/1051-0761(2000)010[1178:CAHIOF]2.0.CO;2>

Wang, T., Hamann, A., Spittlehouse, D., Carroll, C., 2016. Locally Downscaled and Spatially Customizable Climate Data for Historical and Future Periods for North America. PLOS ONE 11, e0156720. <https://doi.org/10.1371/journal.pone.0156720>

Weiss, A., 2001. Topographic position and landforms analysis.

Wright, M.N., Ziegler, A., 2017. Ranger: A fast implementation of random forests for high dimensional data in c++ and r 77. <https://doi.org/10.18637/jss.v077.i01>

Zurell, D., Franklin, J., König, C., Bouchet, P.J., Dormann, C.F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J.J., Leitão, P.J., Park, D.S., Peterson, A.T., Rapacciuolo, G., Schmatz, D.R., Schröder, B., Serra-Diaz, J.M., Thuiller, W., Yates, K.L., Zimmermann, N.E., Merow, C., 2020. A standard protocol for reporting species distribution models. Ecography 43, 1261–1277. <https://doi.org/10.1111/ecog.04960>

# Appendix A: ODMAP

## Overview

Here we describe the SDMs produced herein following the Overview, Data, Model, Assessment, Prediction (ODMAP) protocol for species distribution models ([Zurell et al., 2020](#ref-zurell2020StandardProtocolReporting)). Here, we first provide the Overview for our modeling, while the remaining ODMAP sections are detailed in Table S3.

The objectives of this modelling exercise are to (1) better explain the drivers of aspen’s distribution across the Southern Rocky Mountains, (2) map the area suitable for aspen, and (3) forecast the area suitable for aspen presence in the future under two different climate scenarios.

Table 3: ODMAP protocol information. Details on Data, Model, Assessment, Prediction. For Overview section, please refer to main text.

| ODMAP element | Contents |
| --- | --- |
| Overview |  |
| Authorship | Authors: Sarah J. Hart, Asha Paudel, Maxwell Cook |
|  | Contact email: sarah.hart@colostate.edu |
|  | Title: |
|  | DOI: |
| Model objective | Objective: Inference, Mapping, and Forecasting |
|  | Target outputs: continuous occurrence probabilities, binary maps of potential presence, and maps of potential loss, gain, and stable habitat |
| Taxon | Quaking aspen (Populus tremuloides) |
| Location | Southern Rocky Mountains, US |
| Scale of analysis | Spatial extent (Lon/Lat): Longitude 103.86 ° W - 109.61 ° W, Latitude 35.28 ° N - 47.78 ° N |
|  | Spatial resolution: 90 x 90 m |
|  | Temporal resolution and extent: We modelled the presence of aspen based on remotely sensed maps generated from imagery collected in XXXX |
|  | Type of extent boundary: ecoregion (Southern Rocky Mountains) |
| Biodiversity data overview | Observation type: remotely sensed |
|  | Response type: presence/absence |
| Type of predictors | Climatic, topographic, edaphic |
| Conceptual model / Hypotheses | Based on previous studies, we tested climate, topography and edaphic factors as important environmental predictor variables for aspen presence in the Southern Rocky Mountains. |
| Assumptions | We make the following key assumptions:  (1) aspen is at pseudo-equilibrium with the environment  (2) topography, climate, and soil conditions are the key drivers of aspen's distribution  (3) classification error (estimated at XX) in maps of aspen occurrence were negligible  (4) the relationship fit under current conditions apply to future climate conditions (i.e., no change in key limiting processes) |
| SDM algorithms | Algorithms: SDMs were fit using four different algorithms: generalized linear models (GLMs), generalized additive models (GAMs), gradient boosted trees (GBTs), and random forests (RFs). |
|  | Model complexity: We chose different modelling parameters to optimize each statistical technique. |
|  | Ensembles: We combined the four SDMs to generate an ensemble prediction |
| Model workflow | Prior to model building, all predictor variables were standardized. Model hyperparemeters were then tuned using a spatial cross-validation approach, with the best parameters selected using the area under the reciever operating chracteristic (ROC) curve (AUC). |
| Software | Software: All analyses were conducted using R version 4.3.1 (R Core Team 2023) with the packages XXX, XXX, XXX |
|  | Code availability: All code is publicly available on GitHub (LINK) |
|  | Data availability: Data are available from XXXX |
| Data |  |
| Biodiversity data | Taxon names: Trembling aspen (Populus tremuloides) |
|  | Taxonomic reference system: US Department of Agriculture (USDA) Plant List of Attributes, Names, Taxonomy, and Symbols (PLANTS) Database |
|  | Ecological level: population |
|  | Data source: Aspen presence-absence data at 10x10 m spatial resolution were obtained from Cook et al. (XXXX) |
|  | Sampling design: |
|  | Sample size: The aspen dataset consists of 4,312,302,640 10 x 10 m cells, of which 117,140,964 recorded the presence of aspen. |
|  | Ecoregion mask: We clipped all data to the Southern Rocky Mountain Ecoregion using data from the EPA's (2013) Level III Ecoregions of the Conterminous United States product. |
|  | Scaling: We aggregated the aspen presence-absence data to a 90 x 90 m cell size. |
|  | Data filtering: |
|  | Absence data: The Cook et al. (XXXX) map consists of both presence and absence data. |
| Data partitioning | To reduce computation time, we randomly selected 10,000 cells for model building and 10,000 cells model testing. To reduce the potential effects of spatial autocorrelation, all sample points were separated by a distance of at least 1 km. |
| Environmental data/predictor varaibles | Predictor variables:  (1) Topography: topographic position index, heat load index (HLI)  (2) Climate: 35 climate variables (see Table XX)  (3) Soils: percent clay, percent soil organic matter, saturated water content |
|  | Data sources:  (1) Topography: 3DEP DEM (USGS XXX).  (2) Climate: AdaptWest Project (2022)  (3) Soils: POLARIS soil properties database (Cheney et al. 2019). |
|  | Spatial resolution and extent of raw data: All data were available for the entire study area.  (1) Topography: The raw resolution of the topographic data was 30 m.  (2) Climate: The raw resolution of the topographic data was 1 km.  (3) Soils: The raw resolution of the topographic data was 30 m. |
|  | Temporal resolution and extent of raw data:  (1) Topography: raw topographic data were collected over the period 2009-2023  (2) Climate: mean monthly climate data the period 1981-2010  (3) Soils: integrate data collected over the 1899 to 2019 period as part of the National Cooperative Soil Survey |
|  | Geographic projection of raw data:  (1) Topography: NAD83(HARN) / Conus Albers (EPSG:5071)  (2) Climate: Lambert Azimuthal Equal Area (EPSG:9820)  (3) Soils: WGS 84 (EPSG:4326) |
|  | Data processing: (1) Topography: We cacluated the Heat Load Index (HLI) and topographic position index (TPI) using the spatialeco package (REF). TPI was calculated using a 3 cell neighborhood (i.e. 90 x 90 m) and a 15 cell neighborhood (i.e. 450 x 450 m). TPI and HLI datasets were then aggregated to 90 m using the mean value and reprojected to Universal Transmercator (UTM) Zone 13N.  (2) Climate: We calculated ADI, GSP, PRATIO, and GSPDD5 following Rehfeldt et al. (2009). All climate varaibles were then downscaled to 250 m resolution using gradient and inverse distance squared (GIDS) interpolation.  (3) Soils: Soil data were aggregated to 90 m using the mean value and reprojected to Universal Transmercator (UTM) Zone 13N. |
| Model |  |
| Variable pre-selection | We initially reduced our set of climate predictor variables by first calcuating pairiwise Spearman's correlations and then fitting univriate random forest models of the presence/absenece using raw climate data (i.e., 1 km resoltuion not the 250 m downscaled product). When |r|>0.75, we retained the variable with greatest contribution to AUC, calcuated from the univariate model. |
| Multicollinearity | Using the downscaled climate varibles, we reduced multicollinearity in our predictor dataset by calculating variable inflation factors (VIF). When VIF >5, variables were iteratively removed, giving preference to climate variables (ordered by contribution to AUC from the univariate RF model) using the spatialRF package (Benito 2022). |
| Model settings | (1) The GLM was fit using a logit link function and a binomial error distribution. For all variables, we included both linear and quadratic effects. Models were fit using a Lasso regularization approach. The only tuned the penalty factor. Based on the higest AUC, the penalty factor was set to XX. We did not explore any interaction terms.  (2) The GAM was was fit using a logit link function and a binomial error distribution. For all variables, we used thin plate regression splines as the smooth basis and set bias demensions term (k) to the default value of 10. To prevent overfitted, we allowed We tuned the moothness adjust Models were fit using restricted maximum likelihood (REML). We did not explore any interaction terms.  (3) For the RF model, we tuned the minimum number of data points in a node that is required for the node to be split further (min\_n) and the number of variables to try at each split (mtry). Based on highest AUC, min\_n = XXX an//d mtry = XXX.  (3) GBT: For the GBT model, we tuned the minimum number of data points in a node that is required for the node to be split further (min\_n), the reduction in the loss function required to split further (loss\_reduction), the learning rate (learn\_rate) and the maximum depth of the tree (tree\_depth). Based on highest AUC, min\_n = XXX, loss\_reduction = XXX , learn\_rate = XXX , and tree\_depth = XXX. |
| Model estimates | Using the R package DALEX (REF), we determined variable importance using a model-angostic permutation-based approach. In this approach, each variable is randomized and then the ROC AUC statistic is compared with ROC AUC for the full model (where data has not been randomized). We evaluated the relationship between aspen presene and each predictor variabele using accumulated local effects (ALE) profiles, which were generated using the ingrediates package (REF). |
| Model averaging / Ensembles | We calculated a weighted probabilty of occurrence from all three presence-absence models. Weights assigned were based on the ROC AUC statistic. |
| Non-independence | We evaulated the potenital effects of spatial autocorrelation on our models' predictive ability using a spatially clustered cross-validation approach. |
| Threshold selection | Binary predictions were derived by maximizing Youden's J statistic, which balances sensitivity and specificty (Youden 1950). |
| Assessment |  |
| Performance statistics | Predictive model performancWe on validation data was assessed using X different performance measures: area under the reciever opperator curve (AUC), sensitivity, specificity, overall accuracy, kappa, F measure, Precision, and Recall |
| Plausibility checks | We checked model plausibility by assessing accumulated local effects plots. |
| Prediction |  |
| Prediction output | For further analyses, we used continuous predictions of occurrence probability, as well as predicted presence-absence, which was obtained by binarising the predicted occurrence probabilities using the TSS-maximization threshold. |
| Uncertainty quantification | We account for algorithmic uncertainty by applying an ensemble approach averaging over three different SDM algorithms. |

# Appendix B

Table 4: Climate variables considered for inclusion in SDMs and modeling notes.

| Variable | Description | Findings from previous research | Variable importance | Modeling notes |
| --- | --- | --- | --- | --- |
| ADI | annual dryness index: (DD5^0.5)/MAP | Higher ADI is expected to limit aspen (Rehfeltd et al. 2009) | 0.47883778 | retain |
| GSPDD5 | growing season precipitation to degree day ratio: (GSP\*DD5)/1000 |  | 0.46377305 | removed - initial screening showed strong correlation (r>0.75) with ADI |
| PRATIO | mean precipitation ratio:  GSP/MAP |  | 0.46267798 | retain |
| CMI | Hogg’s climate moisture index (mm) |  | 0.44972308 | removed - initial screening showed strong correlation (r>0.75) with ADI |
| DD\_18 | degree-days below 18 °C |  | 0.40029233 | removed - initial screening showed strong correlation (r>0.75) with ADI |
| DD5 | degree-days above 5 °C |  | 0.36317860 | removed - initial screening showed strong correlation (r>0.75) with ADI |
| MAP | mean annual precipitation (mm) |  | 0.34758509 | removed - initial screening showed strong correlation (r>0.75) with ADI |
| PAS | mean precipitation as snow (mm) between August in previous year and July in current year |  | 0.34581813 | removed - initial screening showed strong correlation (r>0.75) with ADI |
| PPT\_at | mean autumn precipitation (mm) |  | 0.34303178 | removed - initial screening showed strong correlation (r>0.75) with ADI |
| PPT\_wt | mean winter precipitation (mm) |  | 0.31840937 | removed - initial screening showed strong correlation (r>0.75) with ADI |
| DD1040 | degrees-days above 10 °C and below 40 °C |  | 0.31376071 | removed - initial screening showed strong correlation (r>0.75) with ADI |
| DD\_0 | degree-days below 0 °C |  | 0.31238050 | retain |
| CMD | Hargreaves climatic moisture deficit (mm) |  | 0.29730448 | removed - initial screening showed strong correlation (r>0.75) with ADI |
| PPT\_sp | mean spring precipitation (mm) |  | 0.27903248 | removed - initial screening showed strong correlation (r>0.75) with ADI |
| Eref | Hargreave's reference evapotranspiration (mm) |  | 0.27739411 | removed - initial screening showed strong correlation (r>0.75) with ADI |
| MCMT | mean coldest month temperature (°C) |  | 0.23905474 | removed - initial screening showed strong correlation (r>0.75) with DD\_0 |
| MAT | mean annual temperature |  | 0.23380178 | removed - initial screening showed strong correlation (r>0.75) with ADI |
| DD18 | degree-days above 18 °C |  | 0.23028077 | removed - initial screening showed strong correlation (r>0.75) with ADI |
| GSP | growing season (Apr - Sep) precipitation (mm) |  | 0.22964389 | removed during final multicollinearity testing because of high VIF score |
| Tave\_wt | mean winter temperature (°C) |  | 0.22762307 | removed - initial screening showed strong correlation (r>0.75) with DD\_0 |
| Tave\_sm | mean summer temperature (°C) |  | 0.21988789 | removed - initial screening showed strong correlation (r>0.75) with ADI |
| Tave\_sp | mean spring temperature (°C) |  | 0.21929066 | removed - initial screening showed strong correlation (r>0.75) with ADI |
| Tave\_at | mean autumn temperature (°C) |  | 0.21486364 | removed - initial screening showed strong correlation (r>0.75) with ADI |
| EMT | extreme minimum temperature (°C) |  | 0.20953570 | removed - initial screening showed strong correlation (r>0.75) with DD\_0 |
| TMAX | Mean maximum temperature in warmest month |  | 0.20802864 | removed - initial screening showed strong correlation (r>0.75) with DD\_0 |
| MWMT | mean warmest month temperature (°C) |  | 0.20443081 | removed - initial screening showed strong correlation (r>0.75) with ADI |
| NFFD | mean annual number of frost free days |  | 0.20298256 | removed - initial screening showed strong correlation (r>0.75) with DD\_0 |
| EXT | extreme maximum temperature (°C) |  | 0.20272479 | removed - initial screening showed strong correlation (r>0.75) with ADI |
| bFFP | Julian date on which the frost free period beings |  | 0.19909133 | removed - initial screening showed strong correlation (r>0.75) with DD\_0 |
| FFP | mean annual length of the frost-free period (days) |  | 0.19117756 | removed - initial screening showed strong correlation (r>0.75) with DD\_0 |
| TD | difference between MCMT and MWMT (°C) |  | 0.18253739 | removed during final multicollinearity testing because of high VIF score |
| PPT\_sm | mean summer precipitation (mm) |  | 0.16647272 | removed - initial screening showed strong correlation (r>0.75) with TD |
| eFFP | Julian date on which the frost free period ends |  | 0.16054049 | removed - initial screening showed strong correlation (r>0.75) with DD\_0 |
| MAR | mean annual solar radiation (MJ m‐2 d‐1) |  | 0.12135598 | retain |
| RH | mean annual relative humidity (%) |  | 0.09043655 | retain |

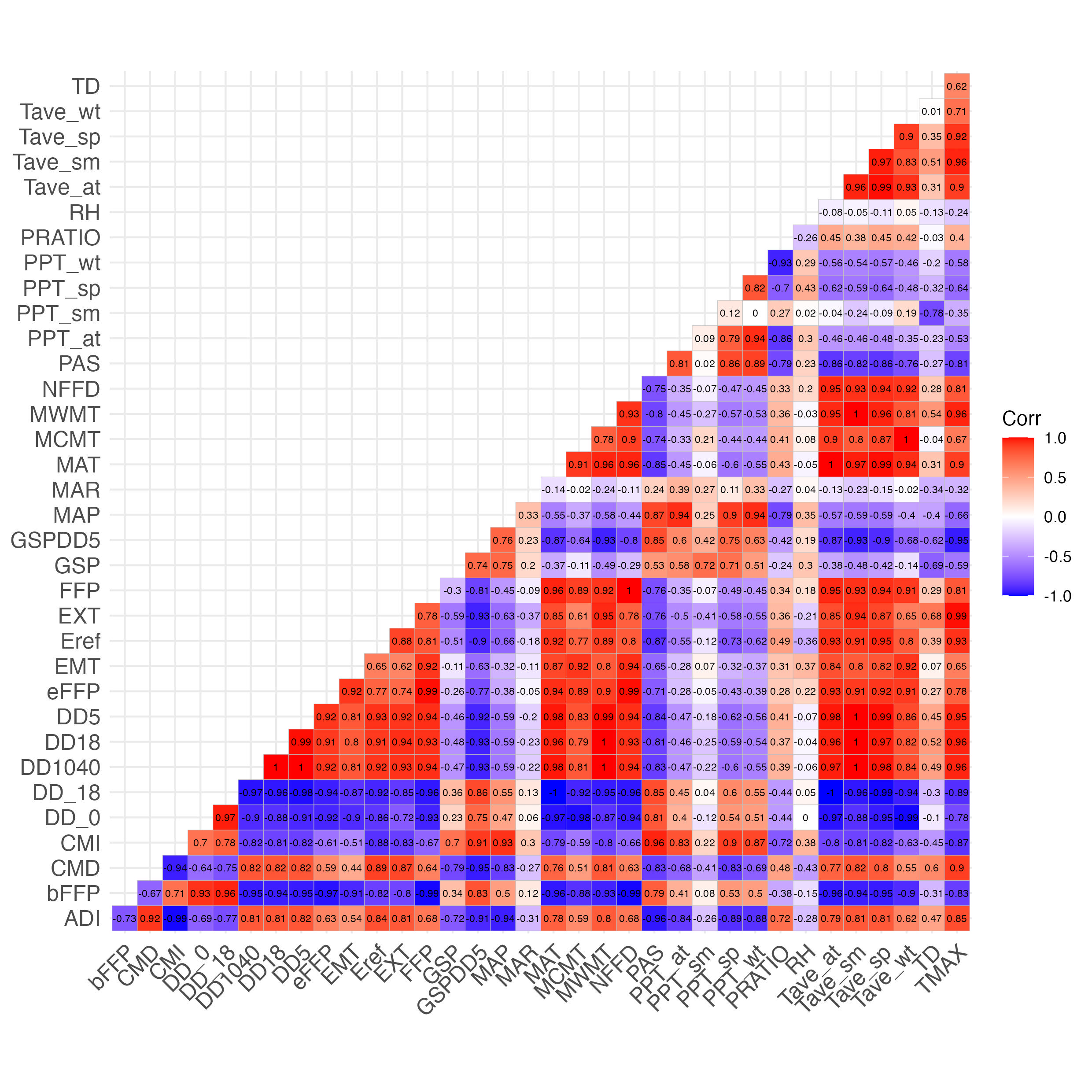


Figure 9: Spearman’s correlation coefficients between pairs of climate predictor variables

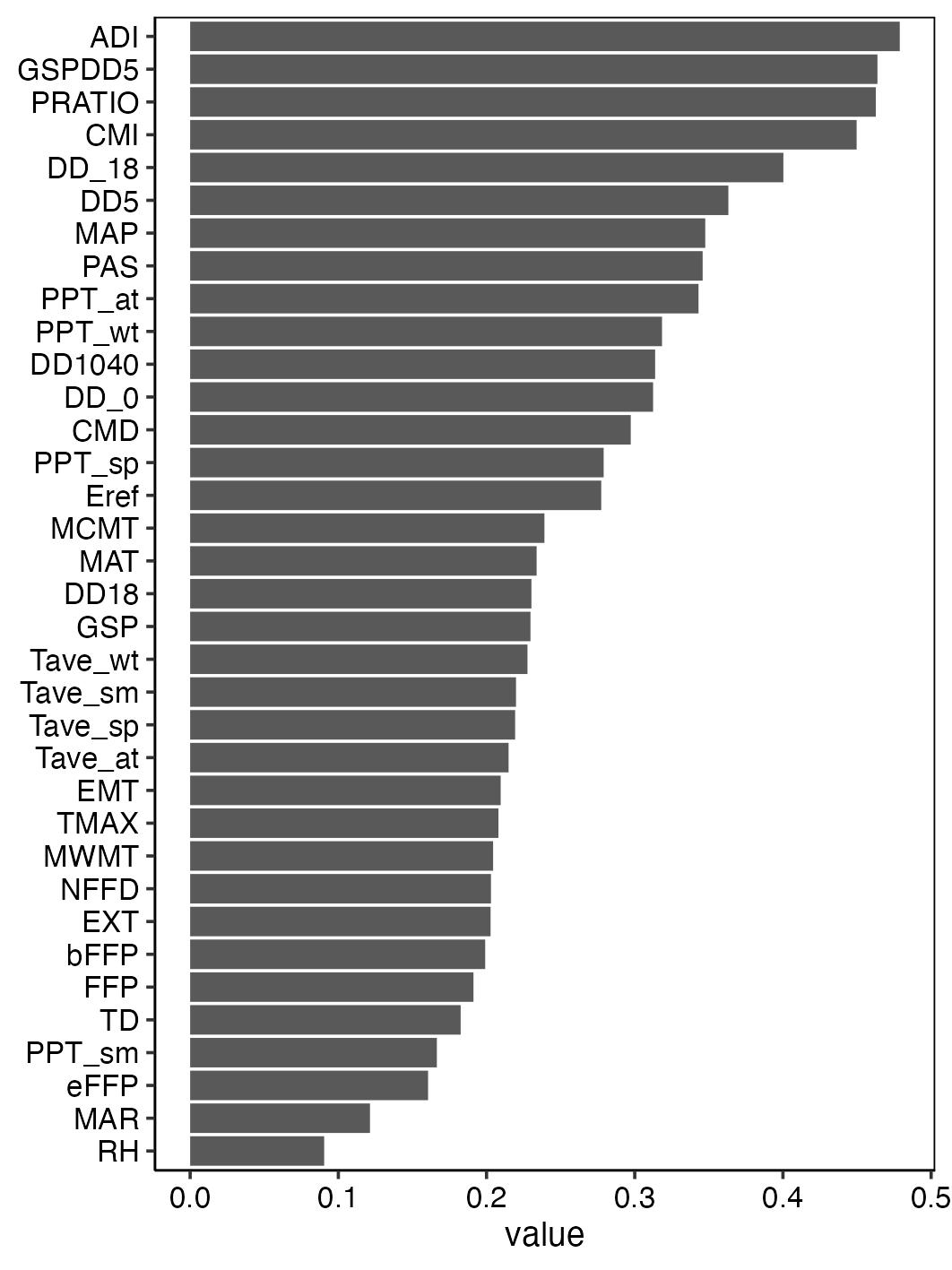


Figure 10: Contribution of climate predictor variables to univariate random forests models

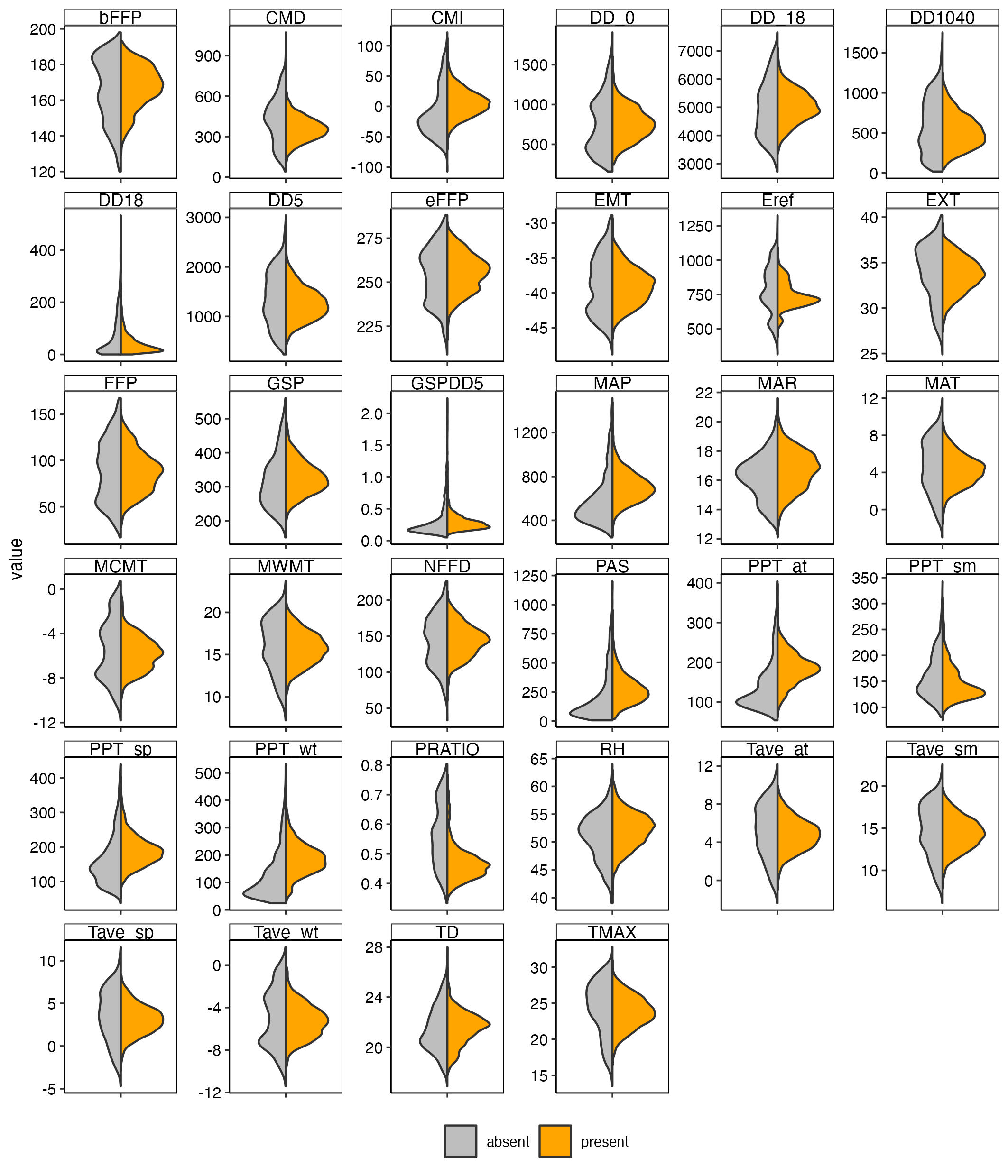


Figure 11: Paired violin plots illustrating the

# Appendix C: Generalized additive modeling

GAMS were fit using restricted maximum likelihood (REML), following recommendations from Pedersen et al.([2018](#X9ce84d2edd409eea4a5bb93b0b43ed50266fe99)). We set the k parameter, which sets t number of basis functions to the default value of 10.

# Appendix D: Model Performance

Table 5: Model performance statistics from spatial cross-validation. Values show the mean ± one standard error.

| Model | Accuracy | F measure | kappa | Precision | Recall | AUC ROC | Sensitivity | Specificity |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| GAM | 0.75 ± 0.02 | 0.75 ± 0.04 | 0.44 ± 0.03 | 0.78 ± 0.04 | 0.73 ± 0.06 | 0.82 ± 0.01 | 0.73 ± 0.06 | 0.72 ± 0.06 |
| GLM | 0.75 ± 0.01 | 0.74 ± 0.04 | 0.44 ± 0.03 | 0.8 ± 0.04 | 0.7 ± 0.06 | 0.8 ± 0.02 | 0.7 ± 0.06 | 0.76 ± 0.04 |
| RF | 0.77 ± 0.02 | 0.78 ± 0.04 | 0.44 ± 0.05 | 0.76 ± 0.03 | 0.81 ± 0.05 | 0.83 ± 0.01 | 0.81 ± 0.05 | 0.61 ± 0.1 |
| GBT | 0.77 ± 0.01 | 0.78 ± 0.04 | 0.44 ± 0.04 | 0.76 ± 0.02 | 0.79 ± 0.05 | 0.83 ± 0.01 | 0.79 ± 0.05 | 0.64 ± 0.1 |

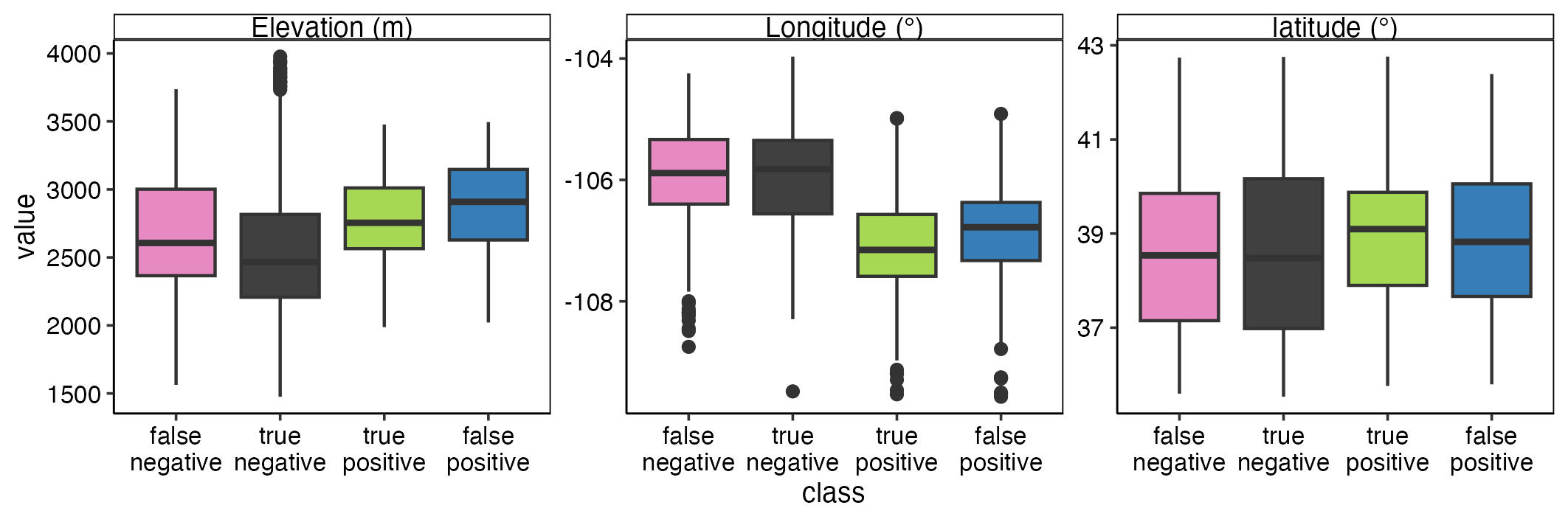


Figure 12: The relationship between geographic position and model performance.

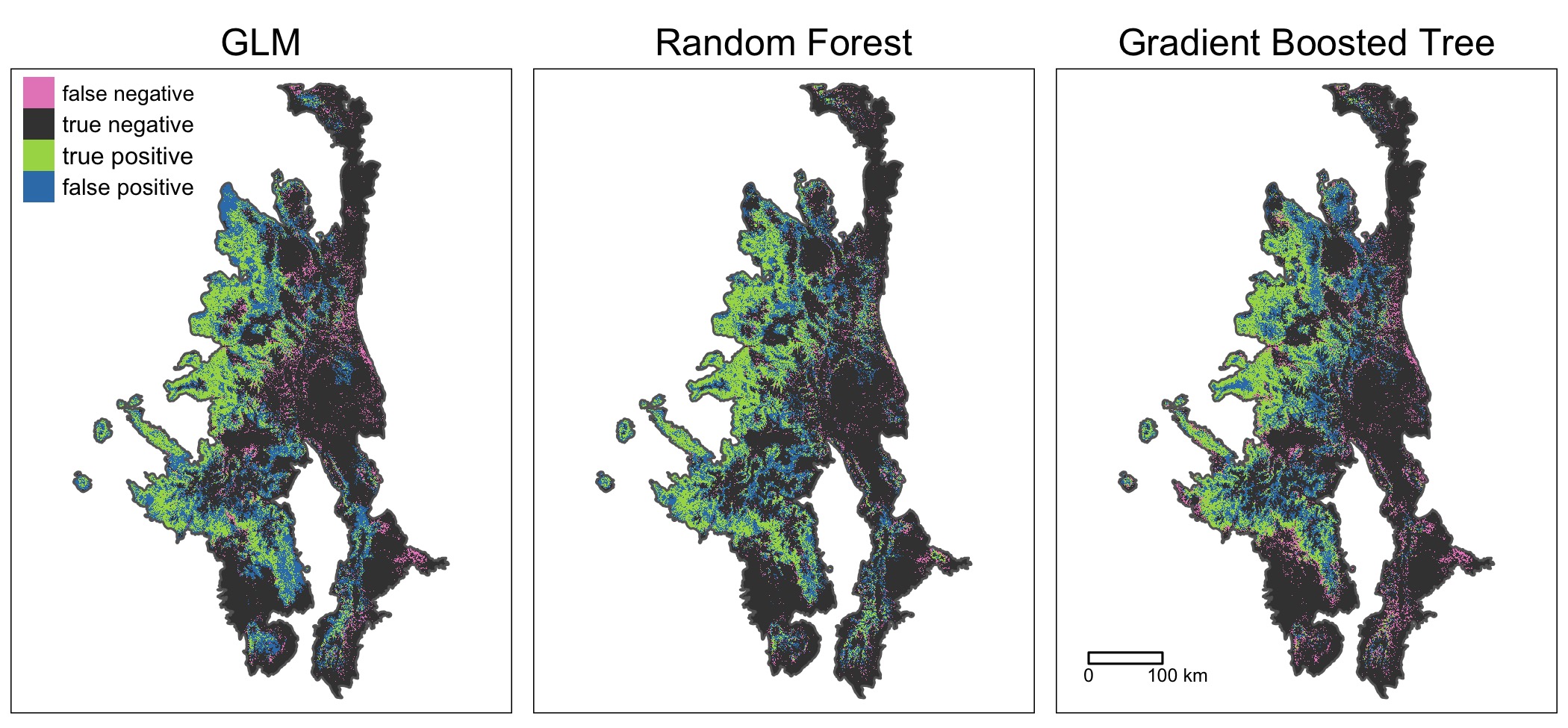


Figure 13: The relationship between geographic position and model performance.

# Appendix D