Eagle Species Distribution Model

Marie Rivers

2022-05-31

This analysis uses the Maxent machine learning technique to model the distribution of bald eagles (Haliaeetus leucocephalus) based on species observations from the Global Biodiversity Information Facility (GBIF.org) and environmental data.

Figure 1: Bald Eagle (Haliaeetus leucocephalus)

Setup

```
# load packages, installing if missing
if (!require(librarian)){
  install.packages("librarian")
  library(librarian)
}
librarian::shelf(
  dismo, dplyr, DT, ggplot2, here, htmltools, leaflet, mapview, purrr, raster, readr, rgbif, rgdal, rJa
select <- dplyr::select # overwrite raster::select
options(readr.show_col_types = FALSE)
```

```
# set random seed for reproducibility
set.seed(80)
```

```
# create directory to store data
dir_data <- here("data/sdm")
dir.create(dir_data, showWarnings = F, recursive = T)
```

Get Species Observations

```
obs_csv <- file.path(dir_data, "obs.csv")
obs_geo <- file.path(dir_data, "obs.geojson")
redo    <- FALSE

if (!file.exists(obs_geo) | redo){
  # get species occurrence data from GBIF with coordinates
  (res <- spocc::occ(
    query = 'Haliaeetus leucocephalus',
    from = 'gbif', has_coords = T, limit = 10000
    ))

  # extract data frame from result
  df <- res$gbif$data[[1]]
  readr::write_csv(df, obs_csv)

  # convert to points of observation from lon/lat columns in data frame
  obs <- df %>%
    sf::st_as_sf(
      coords = c("longitude", "latitude"),
      crs = st_crs(4326)) %>%
    select(prov, key, issues, basisOfRecord, occurrenceStatus, eventDate, isInCluster, lifeStage, locali
  sf::write_sf(obs, obs_geo, delete_dsn=T)
}
obs <- sf::read_sf(obs_geo)
nrow(obs) # number of rows
```

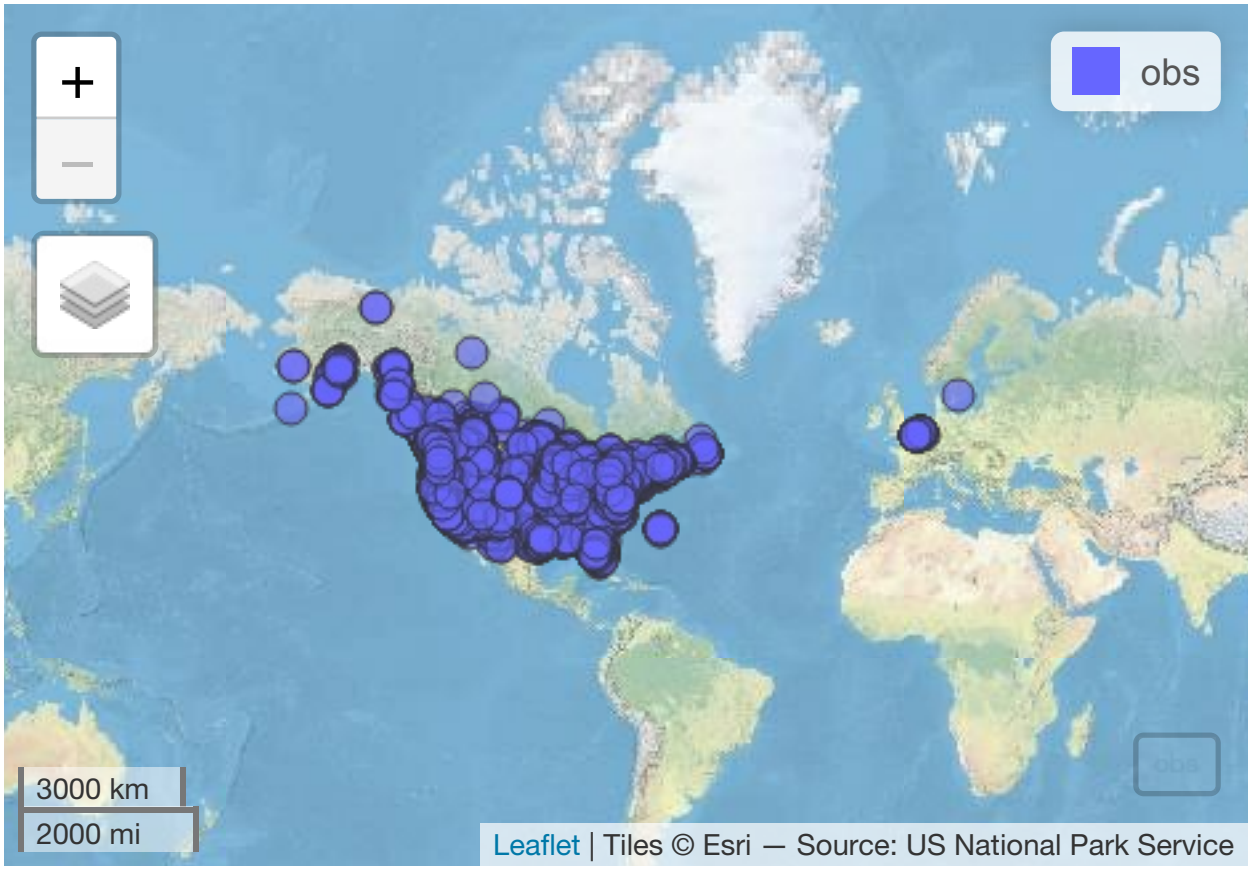
```
## [1] 10000
```

```
# check for observations with duplicate geometries
duplicates <- sum(duplicated(obs$geometry))
```

This model limited the number of observations to 10,000. The `unique()` function was used to check for odd observations. For all observations, the `basisOfrecord` was "human_observation". The `issues` and `occurrenceRemarks` fields didn't not have any concerning entries. The `occurrenceStatus` was "present" for all observations. Observations with duplicate geometries (824) were kept because these records likely indicate mated pairs. While eagles can be found in a variety of habitats, they only build nests in the most desirable locations and therefor these important observations were retained for the analysis.

This map shows the distribution of bald eagle observations.

```
# show points on map
mapview::mapview(obs, map.types = "Esri.WorldPhysical")
```



Get Environmental Data

```
dir_env <- file.path(dir_data, "env")

# set a default data directory
options(sdmpredictors_datadir = dir_env)

# choosing terrestrial
env_datasets <- sdmpredictors::list_datasets(terrestrial = TRUE, marine = FALSE)

# show table of datasets
env_datasets %>%
  select(dataset_code, description, citation) %>%
  DT::datatable()
```

Show 10 entries Search:

	dataset_code	description	citation
1	WorldClim	WorldClim is a set of global climate layers (climate grids). Note that all data has been transformed back to real values, so there is no need to e.g. divide temperature layers by 10.	Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones and A. Jarvis, 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.
4	ENVIREM	The ENVIREM dataset is a set of 16 climatic and 2 topographic variables that can be used in modeling species' distributions. The strengths of this dataset include their close ties to ecological processes, and their availability at a global scale, at several spatial resolutions, and for several time periods. The underlying temperature and precipitation data that went into their construction comes from the WorldClim dataset (www.worldclim.org), and the solar radiation data comes from the Consortium for Spatial Information (www.cgiar-csi.org). The data are compatible with and expand the set of variables from WorldClim v1.4 (www.worldclim.org).	Title, P.O., Bemmels, J.B. 2017. ENVIREM: An expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. Ecography doi: 10.1111/ecog.02880.
5	Freshwater	The dataset consists of near-global, spatially continuous, and freshwater-specific environmental variables in a standardized 1km grid. We delineated the sub-catchment for each grid cell along the HydroSHEDS river network and summarized the upstream environment (climate, topography, land cover, surface geology and soil) to each grid cell using various metrics (average, minimum, maximum, range, sum, inverse distance-weighted average and sum). All variables were subsequently averaged across single lakes and reservoirs of the Global lakes and Wetlands Database that are connected to the river network. Monthly climate variables were summarized into 19 long-term climatic variables following the <d2>bioclim<d3> framework.	Domisch, S., Amatulli, G., and Jetz, W. (2015) Near-global freshwater-specific environmental variables for biodiversity analyses in 1 km resolution. Scientific Data 2:150073 doi: 10.1038/sdata.2015.73

Showing 1 to 3 of 3 entries Previous 1 Next

```
# choose datasets for a vector
env_datasets_vec <- c("WorldClim", "ENVIREM")

# get layers
env_layers <- sdmpredictors::list_layers(env_datasets_vec)
DT::datatable(env_layers)
```

Show 10 entries Search:

	dataset_code	layer_code	name	description	terrestrial	marine	freshwater	cellsize_equalarea	cellsize_lonlat	units	primary_type	primary_spatial_resolution	primary_source	start_year	start_month	start_day	end_year	end_month	end_day	derivation	month	is_surface	version	layer_url
1	WorldClim	WC_alt	Altitude	Altitude	true	false	false	7000	0.083333333	meters	Satellite (SRTM), GTOPO30	3 arcsecond, 30 arsecond	SRTM: CGIAR-CSI, reference: Jarvis, A., H.I. Reuter, A. Nelson, E. Guevara, 2008, Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database (http://srtm.csi.cgiar.org). GTOPO30: url: https://lta.cr.usgs.gov/GTOPO30	1950	1	1	2000	1	1	"		true	1	https://www.lifewatch.be/sdmpredictors/WC_alt_lonlat.tif
2	WorldClim	WC_bio1	Annual mean temperature	Annual mean temperature	true	false	false	7000	0.083333333	Celsius	in situ measurement	"	Weather station data from multiple sources, see Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones and A. Jarvis, 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.	1950	1	1	2000	1	1	mean, thin-plate smoothing splines		true	1	https://www.lifewatch.be/sdmpredictors/WC_bio1_lonlat.tif
3	WorldClim	WC_bio2	Mean diurnal temperature range	Mean of the monthly (maximum temperature - minimum temperature)	true	false	false	7000	0.083333333	Celsius	in situ measurement	"	Weather station data from multiple sources, see Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones and A. Jarvis, 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.	1950	1	1	2000	1	1	mean range, thin-plate smoothing splines		true	1	https://www.lifewatch.be/sdmpredictors/WC_bio2_lonlat.tif
4	WorldClim	WC_bio3	Isothermality	Mean diurnal temperature range (bio2) / Annual temperature range (bio7)	true	false	false	7000	0.083333333	Celsius	in situ measurement	"	Weather station data from multiple sources, see Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones and A. Jarvis, 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.	1950	1	1	2000	1	1	"		true	1	https://www.lifewatch.be/sdmpredictors/WC_bio3_lonlat.tif
5	WorldClim	WC_bio4	Temperature seasonality	Standard deviation of the annual mean temperature	true	false	false	7000	0.083333333	Celsius	in situ measurement	"	Weather station data from multiple sources, see Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones and A. Jarvis, 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.	1950	1	1	2000	1	1	standard deviation, thin-plate smoothing splines		true	1	https://www.lifewatch.be/sdmpredictors/WC_bio4_lonlat.tif
6	WorldClim	WC_bio5	Maximum temperature	Maximum temperature of the warmest month	true	false	false	7000	0.083333333	Celsius	in situ measurement	"	Weather station data from multiple sources, see Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones and A. Jarvis, 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.	1950	1	1	2000	1	1	maximum, thin-plate smoothing splines		true	1	https://www.lifewatch.be/sdmpredictors/WC_bio5_lonlat.tif
7	WorldClim	WC_bio6	Minimum temperature	Minimum temperature of the coldest month	true	false	false	7000	0.083333333	Celsius	in situ measurement	"	Weather station data from multiple sources, see Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones and A. Jarvis, 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.	1950	1	1	2000	1	1	minimum, thin-plate smoothing splines		true	1	https://www.lifewatch.be/sdmpredictors/WC_bio6_lonlat.tif
8	WorldClim	WC_bio7	Annual temperature range	Maximum temperature (bio5) minimum temperature (bio6)	true	false	false	7000	0.083333333	Celsius	in situ measurement	"	Weather station data from multiple sources, see Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones and A. Jarvis, 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.	1950	1	1	2000	1	1	range, thin-plate smoothing splines		true	1	https://www.lifewatch.be/sdmpredictors/WC_bio7_lonlat.tif
9	WorldClim	WC_bio8	Mean temperature of wettest quarter	Mean temperature of wettest quarter	true	false	false	7000	0.083333333	Celsius	in situ measurement	"	Weather station data from multiple sources, see Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones and A. Jarvis, 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.	1950	1	1	2000	1	1	mean, thin-plate smoothing splines		true	1	https://www.lifewatch.be/sdmpredictors/WC_bio8_lonlat.tif
10	WorldClim	WC_bio9	Mean temperature of driest quarter	Mean temperature of driest quarter	true	false	false	7000	0.083333333	Celsius	in situ measurement	"	Weather station data from multiple sources, see Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones and A. Jarvis, 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.	1950	1	1	2000	1	1	mean, thin-plate smoothing splines		true	1	https://www.lifewatch.be/sdmpredictors/WC_bio9_lonlat.tif

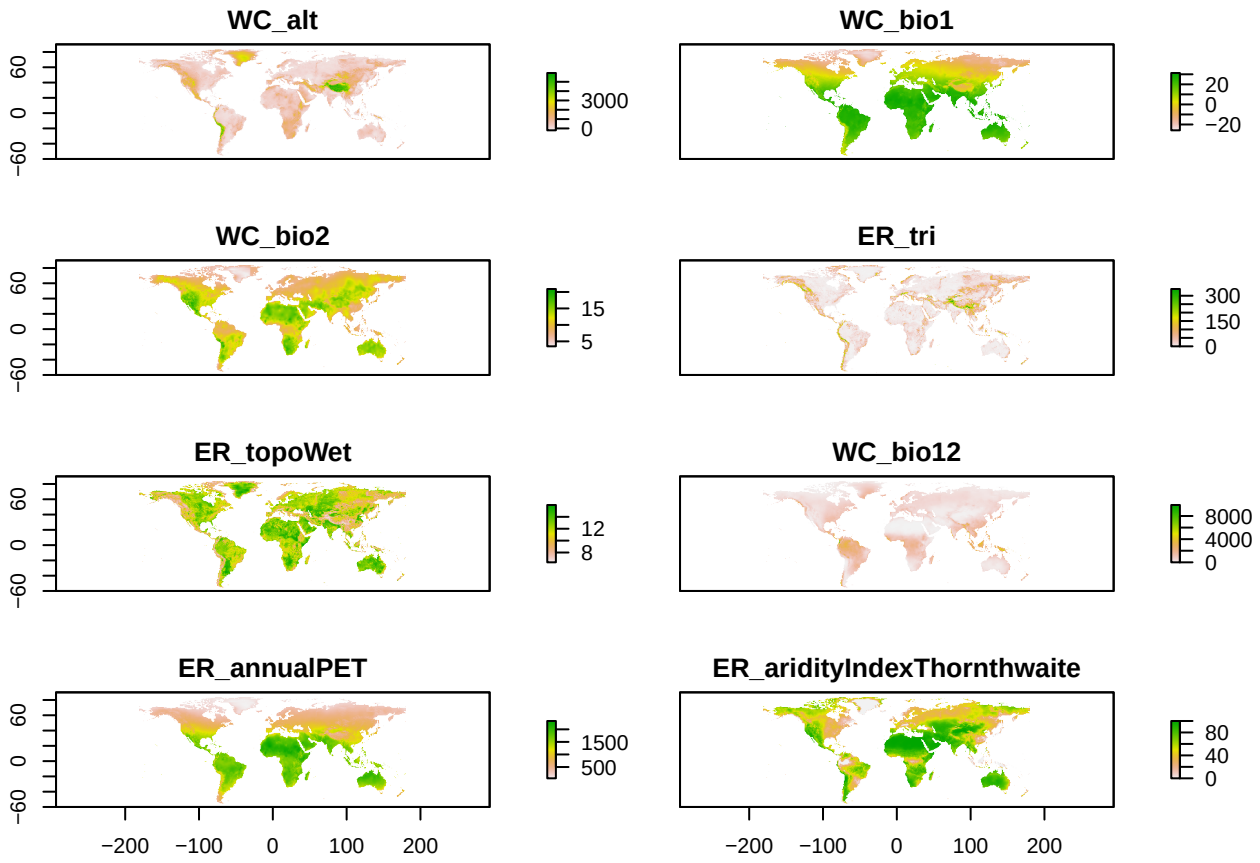
Showing 1 to 10 of 86 entries Previous 1 2 3 4 5 … 9 Next

Based on the literature, bald eagles prefer habits near wetlands and open bodies of water with abundance fish such as seacoasts, rivers, lakes, and marshes. Eagles also prefer trees near water, particularly old growth and mature stands of coniferous or hardwood trees. Eagles can live in a wide range of temperatures.

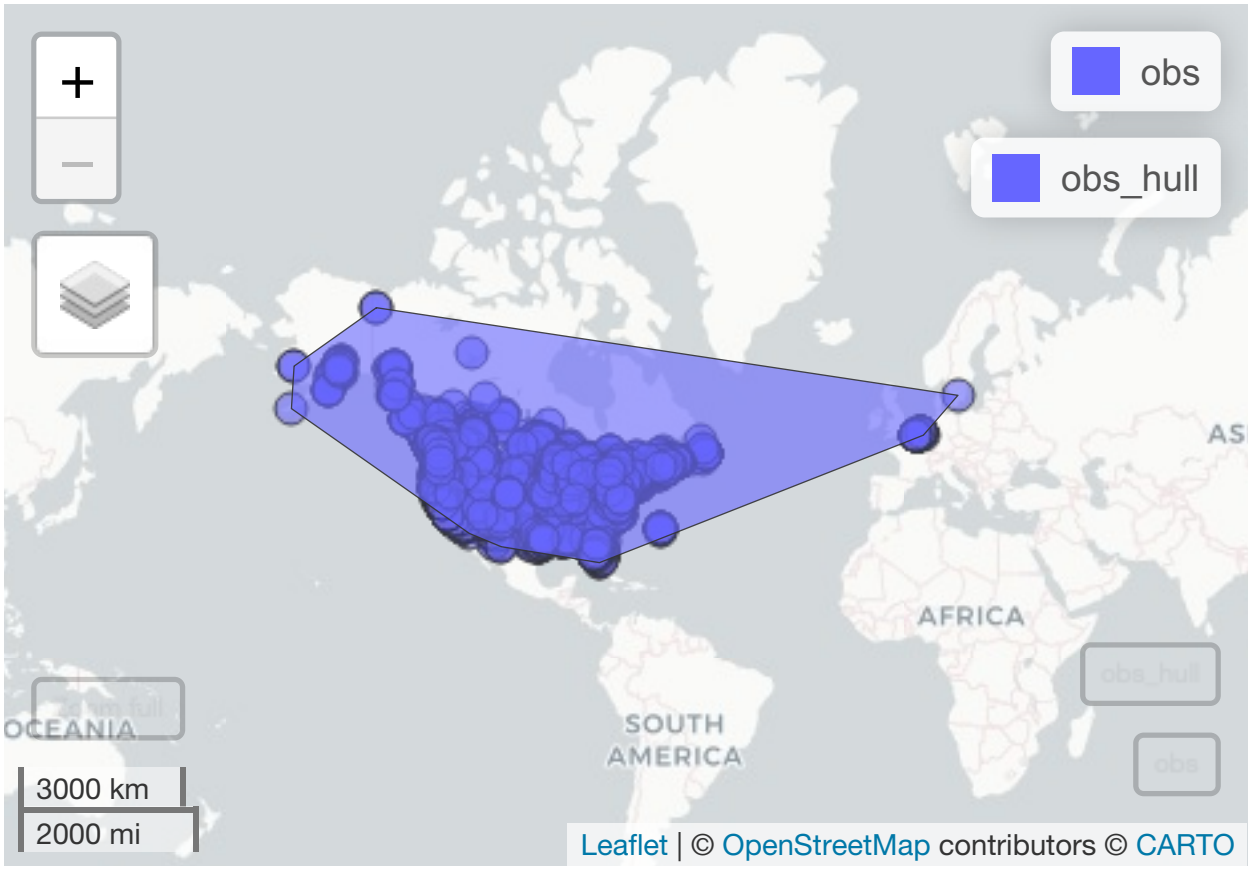
The environmental predictors selected for this analysis included: altitude, annual mean temperature, mean diurnal temperature range, terrain roughness index, topographic wetness, annual precipitation, annual potential evapotranspiration, and the Thornthwaite aridity index which is an index of the degree of water deficit below water need.

Other potentially useful predictors that were not available for the WorldClim or ENVIREM data sets include forested area and proximity to low/medium/high density populations of humans.

```
# choosen layers after consulting literature
env_layers_vec <- c("WC_alt", "WC_bio1", "WC_bio2", "ER_tri", "ER_topoWet", "WC_bio12", "ER_annualPET",

# get layers
env_stack <- load_layers(env_layers_vec)
# plot layers
plot(env_stack, nc=2)
```

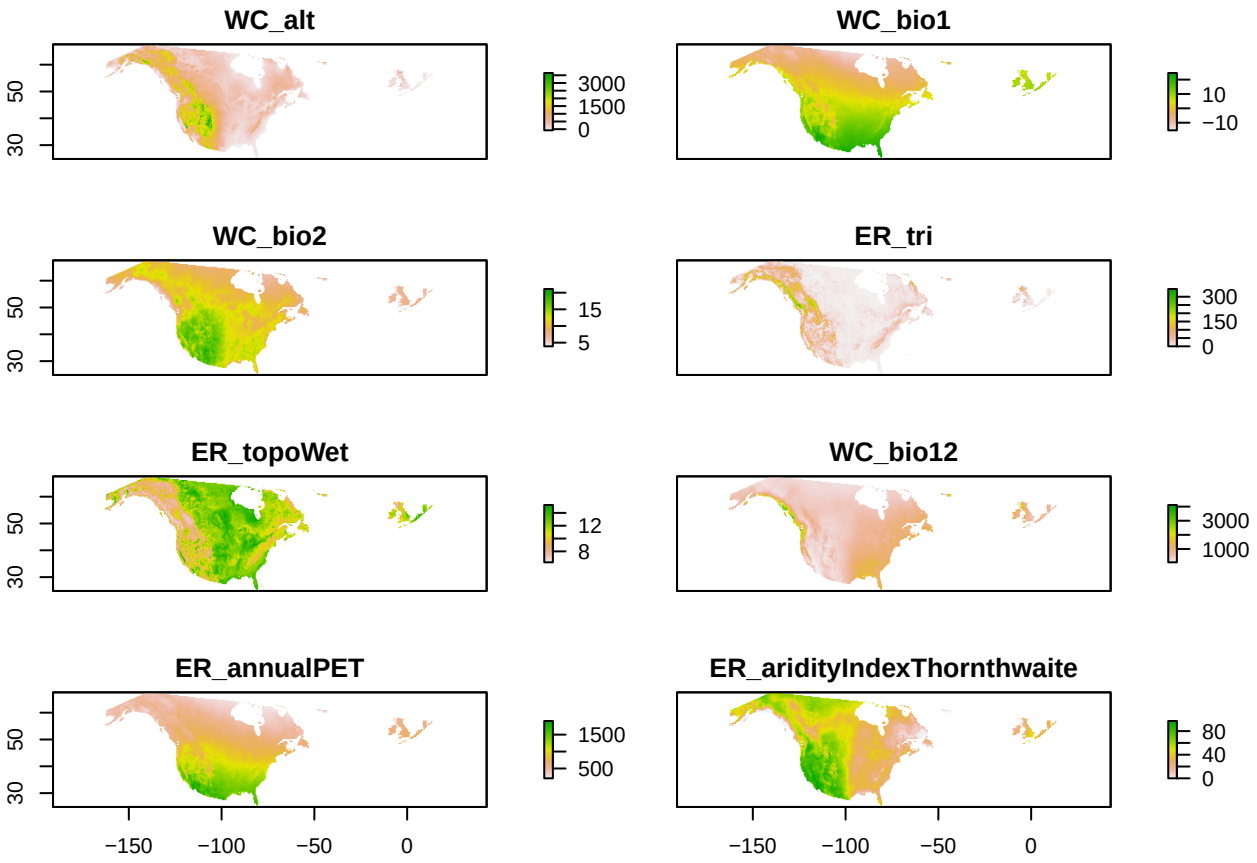



```
# crop the environmental rasters to a reasonable study area around the species observations
obs_hull_geo  <- file.path(dir_data, "obs_hull.geojson")
env_stack_grd <- file.path(dir_data, "env_stack.grd")

if (!file.exists(obs_hull_geo) | redo){
  # make convex hull around points of observation
  obs_hull <- sf::st_convex_hull(st_union(obs))

  # save obs hull
  write_sf(obs_hull, obs_hull_geo)
}
obs_hull <- read_sf(obs_hull_geo)

# show points on map
mapview(
  list(obs, obs_hull))
```



Plots of environmental raster layers clipped to the bald eagle range

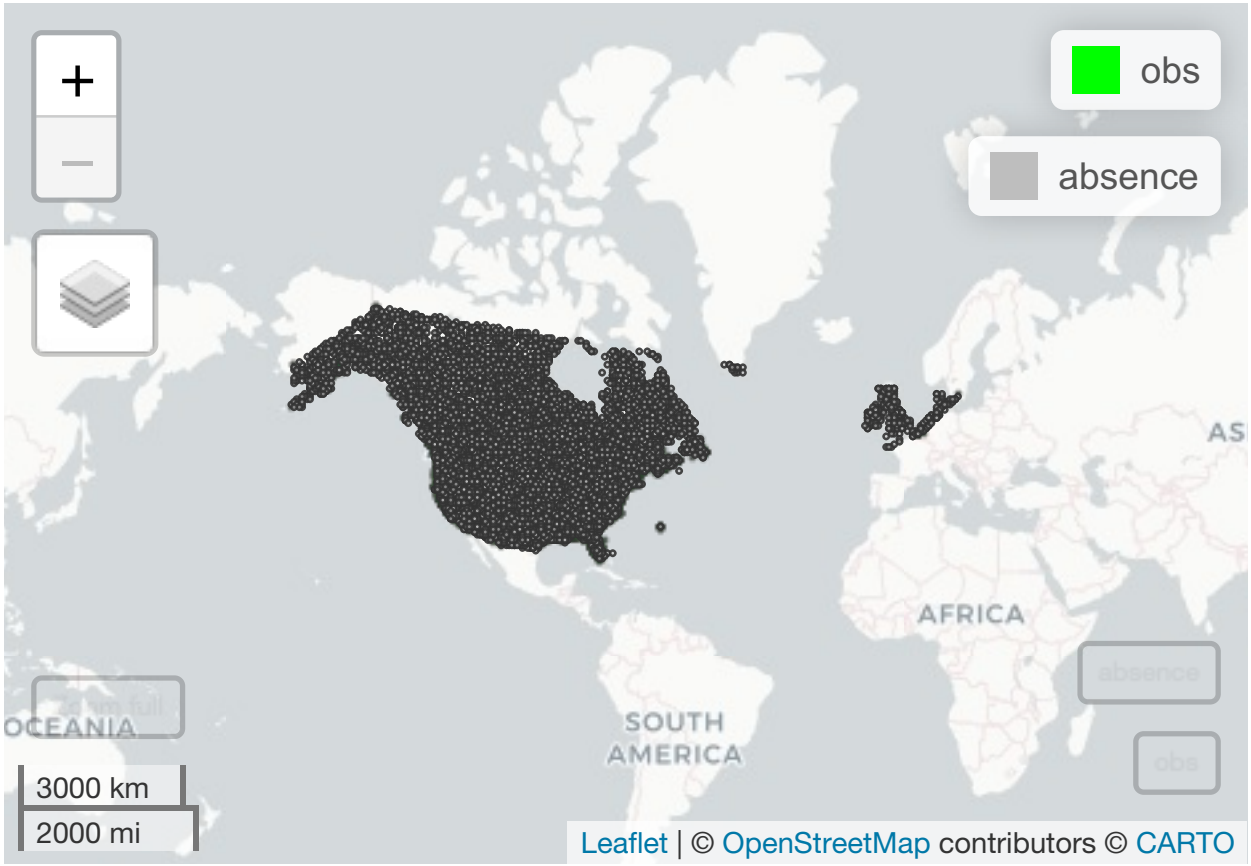
```
if (!file.exists(env_stack_grd) | redo){
  obs_hull_sp <- sf::as_Spatial(obs_hull)
  env_stack <- raster::mask(env_stack, obs_hull_sp) %>%
    raster::crop(extent(obs_hull_sp))
  writeRaster(env_stack, env_stack_grd, overwrite=T)
}
env_stack <- stack(env_stack_grd)

plot(env_stack, nc=2)
```



Pseudo-Absence

```
absence_geo <- file.path(dir_data, "absence.geojson")
pts_geo     <- file.path(dir_data, "pts.geojson")
pts_env_csv <- file.path(dir_data, "pts_env.csv")

if (!file.exists(absence_geo) | redo){
  # get raster count of observations
  r_obs <- rasterize(
    sf::as_Spatial(obs), env_stack[[1]], field=1, fun='count')

  # create mask for
  r_mask <- mask(env_stack[[1]] > -Inf, r_obs, inverse=T)

  # generate random points inside mask
  absence <- dismo::randomPoints(r_mask, nrow(obs)) %>%
    as_tibble() %>%
    st_as_sf(coords = c("x", "y"), crs = 4326)

  write_sf(absence, absence_geo, delete_dsn=T)
}
absence <- read_sf(absence_geo)

# show map of presence, ie obs, and absence
```

```
mapview(obs, col.regions = "green", cex = 0.75) +
  mapview(absence, col.regions = "gray", cex = 0.25)
```




```
if (!file.exists(pts_env_csv) | redo){

  # combine presence and absence into single set of labeled points
  pts <- rbind(
    obs %>%
      mutate(
        present = 1) %>%
      select(present, key),
    absence %>%
      mutate(
        present = 0,
        key     = NA)) %>%
    mutate(
      ID = 1:n()) %>%
    relocate(ID)
  write_sf(pts, pts_geo, delete_dsn=T)

  # extract raster values for points
  pts_env <- raster::extract(env_stack, as_Spatial(pts), df=TRUE) %>%
    tibble() %>%
    # join present and geometry columns to raster value results for points
    left_join(
```

```
      pts %>%
        select(ID, present),
      by = "ID") %>%
    relocate(present, .after = ID) %>%
    # extract lon, lat as single columns
    mutate(
      #present = factor(present),
      lon = st_coordinates(geometry)[,1],
      lat = st_coordinates(geometry)[,2]) %>%
    select(-geometry)
  write_csv(pts_env, pts_env_csv)
}
pts_env <- read_csv(pts_env_csv)

pts_env %>%
  # show first 10 presence, last 10 absence
  slice(c(1:10, (nrow(pts_env)-9):nrow(pts_env))) %>%
  DT::datatable(
    rownames = F,
    options = list(
      dom = "t",
      pageLength = 20))
```

ID	present	WC_alt	WC_bio1	WC_bio2	ER_tri	ER_topoWet	WC_bio12	ER_annualPET	ER_aridityIndexThornthwaite	lon	lat
1	1	34	9.69999980926514	7.30000019073486	14	10.9300003051758	763	704.640014648438	69.879997253418	-123.3734	48.4644
2	1	1695	10.6999998092651	17.3999996185303	36.189998626709	10.6300001144409	368	1292.06005859375	76.3000030517578	-112.682446	37.221921
3	1	79	11.1999998092651	11.5	11.6499996185303	11.4700002670288	1156	1051	22.6499996185303	-75.036235	40.30296
4	1	125	6.19999980926514	9.39999961853027	35.810001373291	9.55000019073486	1284	765.090026855469	23.3799991607666	-65.345831	44.903986
5	1	253	19.1000003814697	12.3999996185303	7.98999977111816	12.6499996185303	838	1436.14001464844	41.6500015258789	-97.67362	30.720572
6	1	239	7.19999980926514	11.1999998092651	9.53999996185303	12.2600002288818	735	904.330017089844	28.1900005340576	-93.188863	44.874098
7	1	190	9.39999961853027	11.3000001907349	22.7000007629395	10.8199996948242	1219	968.109985351562	27.0400009155273	-73.577427	41.357217
8	1	7	10	8.69999980926514	9.8100004196167	11.7799997329712	1066	779.75	53.8699989318848	-123.052933	49.024492
9	1	4	22.8999996185303	11.3000001907349	0.860000014305115	13.3900003433228	1259	1514.7099609375	45.4500007629395	-82.307236	26.88035
10	1	32	18.7000007629395	12.3999996185303	7.38000011444092	12.3199996948242	1521	1407.68005371094	28.4699993133545	-91.439707	31.352648
19991	0	1202	-1	10.3999996185303	84.9899978637695	8.5600004196167	554	553.539978027344	32.2400016784668	-122.958333333333	56.7083333333333
19992	0	1356	7.69999980926514	16.7999992370605	15.3500003814697	11.1599998474121	316	1117.64001464844	79.8000030517578	-117.125	42.9583333333333
19993	0	21	9.60000038146973	8.80000019073486	20.7099990844727	11.5900001525879	935	773.179992675781	54.560001373291	-122.791666666666	48.875
19994	0	472	-3.79999995231628	10.8999996185303	9.27999973297119	11.9300003051758	866	505.709991455078	15.0299997329712	-64.6249999999999	53.4583333333333
19995	0	524	-1.10000002384186	11.5	25.8199996948242	10.8100004196167	991	618.070007324219	14.9300003051758	-69.4583333333332	49.9583333333333
19996	0	125	17.7999992370605	13.3000001907349	8.25	12.1599998474121	1165	1406.90002441406	40.9199981689453	-83.7916666666665	32.7083333333333
19997	0	236	1.79999995231628	11	16.9599990844727	11.3299999237061	782	689.190002441406	32.9799995422363	-88.1249999999998	48.7083333333333
19998	0	21	8	6.90000009536743	19.8999996185303	10.3000001907349	690	545.969970703125	34.4599990844727	-4.04166666666669	57.7083333333333
19999	0	221	-2.90000009536743	9.89999961853027	8.94999980926514	11.710000038147	735	509.679992675781	26.7700004577637	-76.1249999999999	52.9583333333333
20000	0	986	-7.09999990463257	13.3000001907349	54.939998626709	9.60999965667725	217	524.340026855469	67.2900009155273	-143.291666666666	64.2083333333333

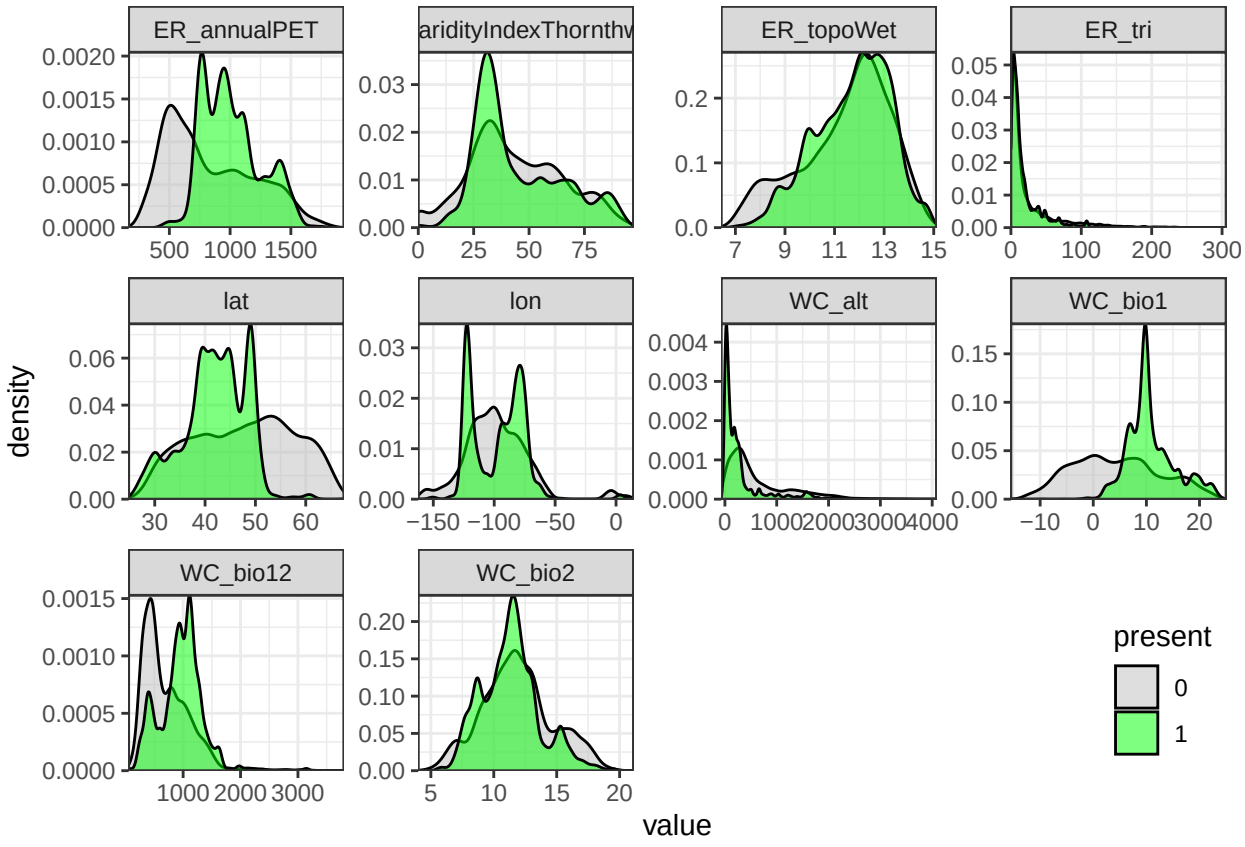
```
# check that all presence and absence points are included
nrow(pts_env)
```

```
## [1] 20000
```

Term Plots

The term plots display predictors and responses. For modeling purposes, predictors are preferred where presence occupies a distinct niche from the background absence points. The term plots are a good way visualize how differentiated presence is from absence for each predictor.

```
pts_env %>%
  select(-ID) %>%
  mutate(
    present = factor(present)) %>%
  pivot_longer(-present) %>%
  ggplot() +
  geom_density(aes(x = value, fill = present)) +
  scale_fill_manual(values = alpha(c("gray", "green"), 0.5)) +
  scale_x_continuous(expand=c(0,0)) +
  scale_y_continuous(expand=c(0,0)) +
  theme_bw() +
  facet_wrap(~name, scales = "free") +
  theme(
    legend.position = c(1, 0),
    legend.justification = c(1, 0))
```



Based on the results of Term Plots, topographic wetness (ER_topoWet) and terrain roughness index (ER_tri) are similarly distributed for presence and absence points and are therefore likely not strong predictors for a bald eagle species distribution model. The term plots suggest that annual mean temperature (WC_bio1), annual precipitation (WC_bio12), and annual potential evapotranspiration (ER_annualPET) could be useful predictors of bald eagle species distribution.

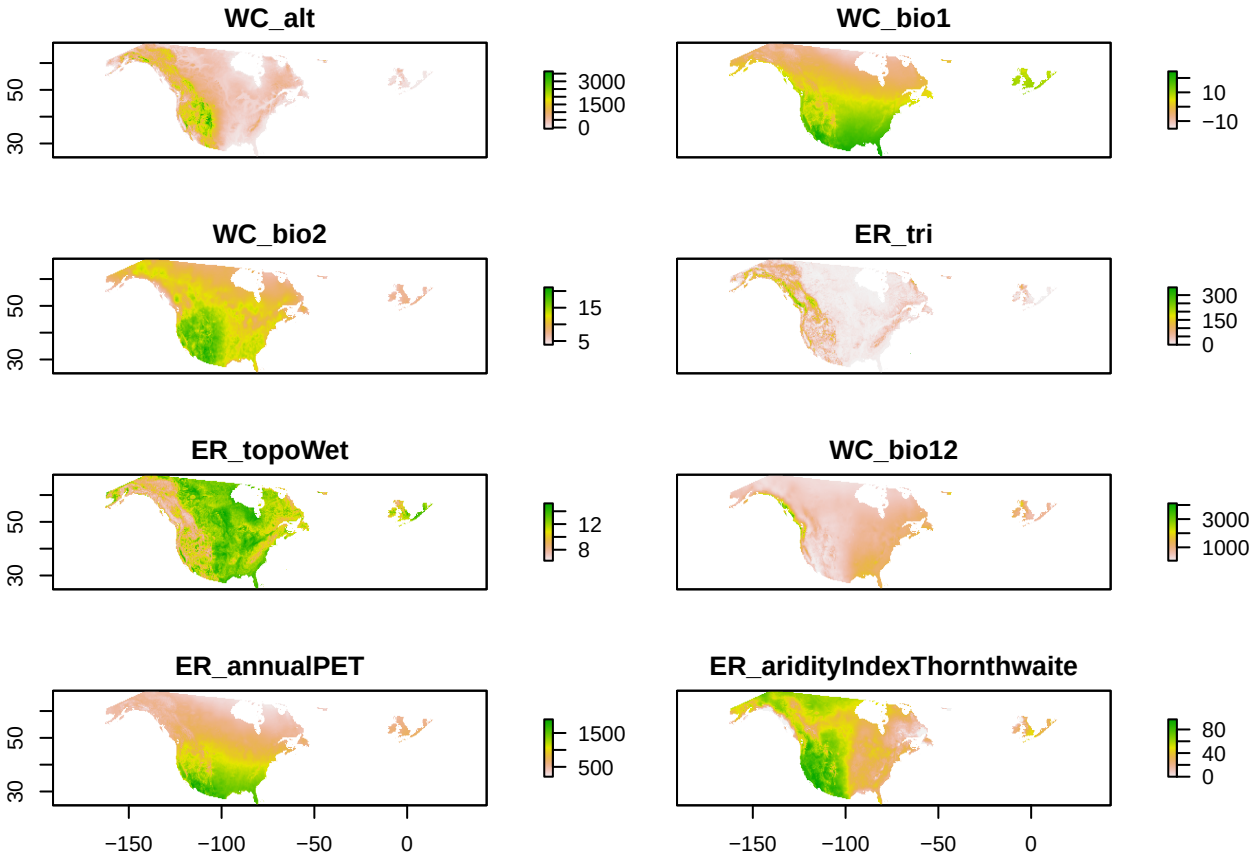
Maxent (Maximum Entropy)

Maxent is a commonly used species distribution model that performs well with few input data points and only requires presence points (background 'absence' points generated during the analysis). Since this example only has presence points, the background is sampled for comparison.

```
# load extra packages
librarian::shelf(
  maptools, sf)

mdl_maxent_rds <- file.path(dir_data, "mdl_maxent.rds")

# show version of maxent
if (!interactive())
  maxent()
```

```
## This is MaxEnt version 3.4.3
```

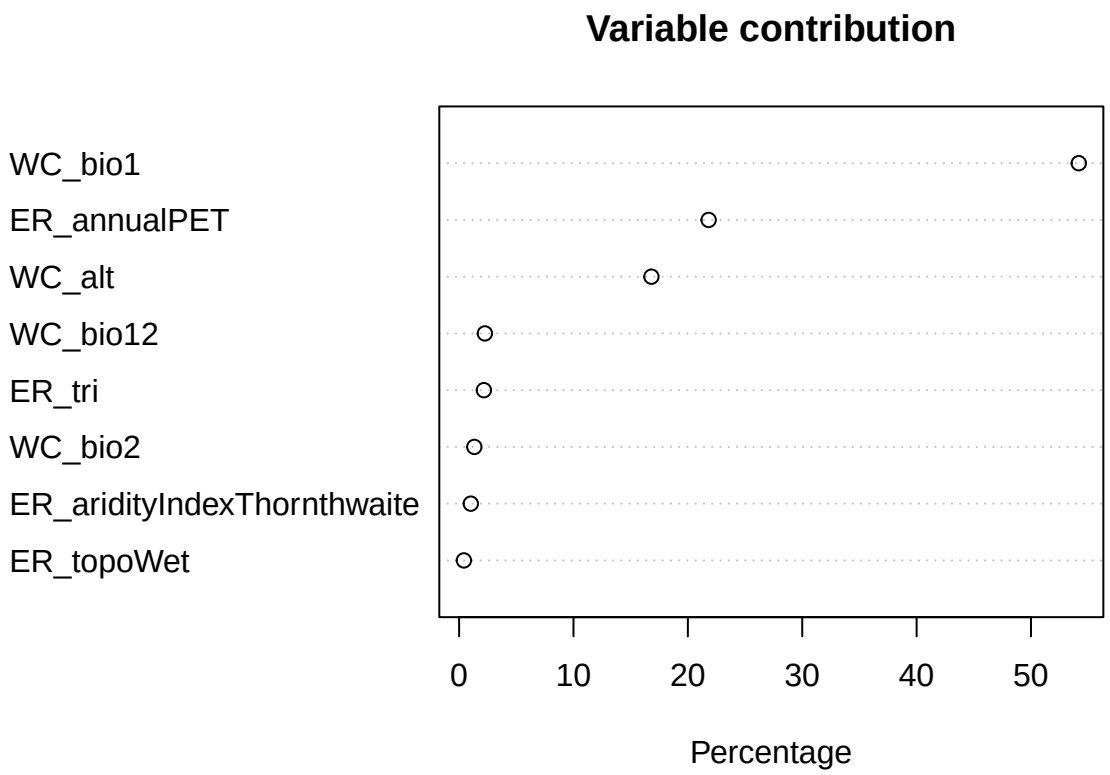
```
env_stack_grd <- file.path(dir_data, "env_stack.grd")
env_stack <- stack(env_stack_grd)
plot(env_stack, nc=2)
```



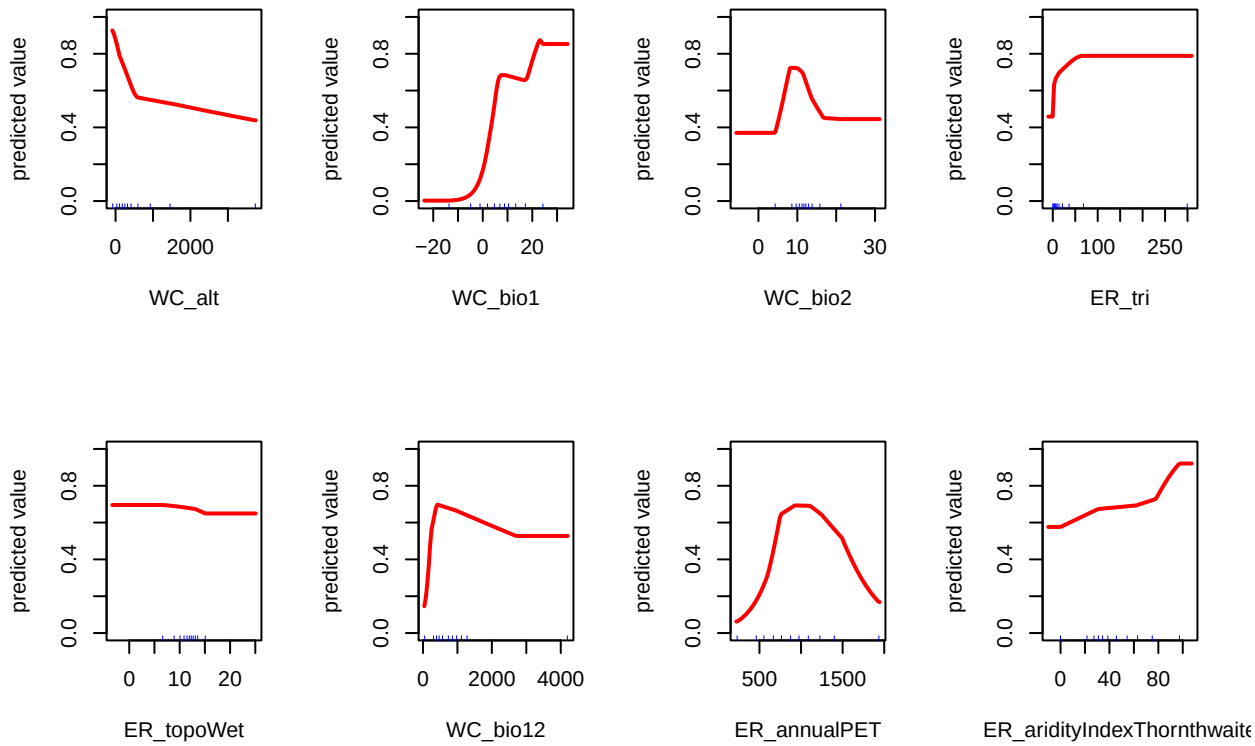

```
# get the presence-only observation points (maxent extracts raster values for you)
obs_geo <- file.path(dir_data, "obs.geojson")
obs_sp <- read_sf(obs_geo) %>%
  sf::as_Spatial() # maxent prefers sp::SpatialPoints over newer sf::sf class
```

```r
# fit a maxent entropy model
if (!file.exists(mdl_maxent_rds)){
  mdl_maxent <- maxent(env_stack, obs_sp)
  readr::write_rds(mdl_maxent, mdl_maxent_rds)
}
mdl_maxent <- read_rds(mdl_maxent_rds)
```

```r
# plot variable contributions per predictor
plot(mdl_maxent)
```

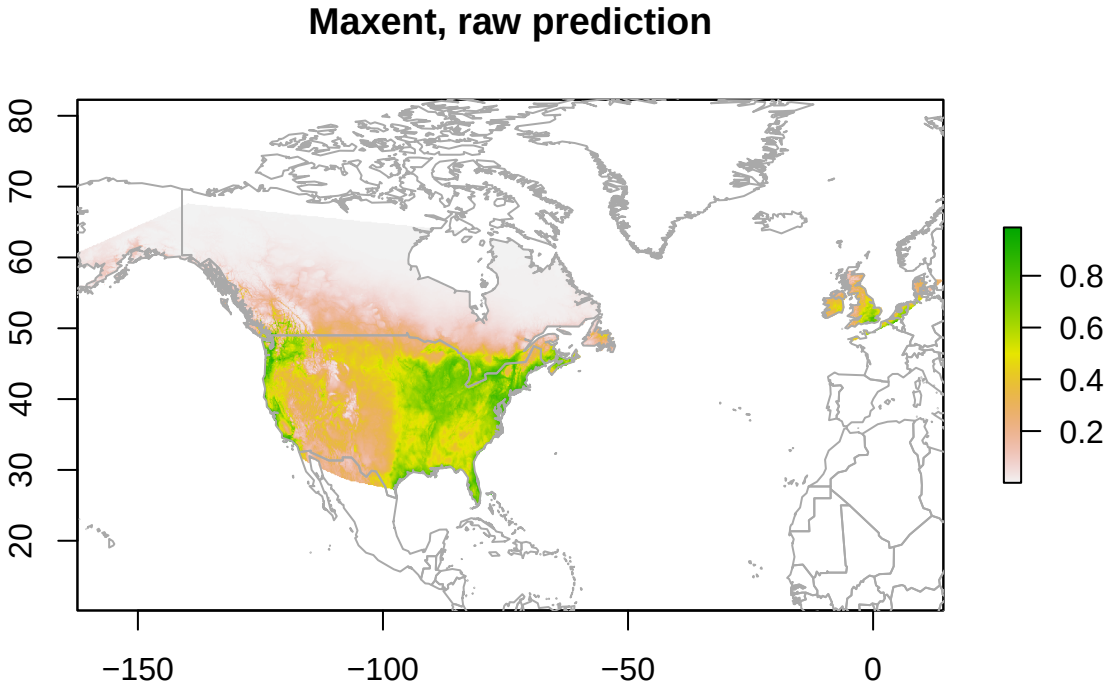



```r
# plot term plots
response(mdl_maxent)
```



WC_bio1 contributes most to the Maxent predictions.

```
# predict
y_predict_maxent <- predict(env_stack, mdl_maxent) #, ext=ext, progress='')

plot(y_predict_maxent, main='Maxent, raw prediction')
data(wrld_simpl, package = "maptools")
plot(wrld_simpl, add=TRUE, border='dark grey')
```



```
# paths
mdl_maxv_rds <- file.path(dir_data, "mdl_maxent_vif.rds")
```

Based on the results of the Maxent model, mean annual temperature (WC_bio1) contributes most towards predicting presences, followed by altitude (WC_alt).

Evaluate: Model Performance

```
librarian::shelf(usdm) # uncertainty analysis for species distribution models: vifcor()
```

```
# read points of observation: presence (1) and absence (0)
pts <- read_sf(pts_geo)

# read raster stack of environment
env_stack < raster::stack(env_stack_grd)
```

```
## class      : RasterBrick
## dimensions : 513, 2120, 1087560, 8  (nrow, ncol, ncell, nlayers)
## resolution : 0.08333333, 0.08333333  (x, y)
## extent     : -162.3333, 14.33333, 24.83333, 67.58333  (xmin, xmax, ymin, ymax)
## crs        : +proj=longlat +datum=WGS84 +no_defs
## source     : memory
## names      : WC_alt, WC_bio1, WC_bio2, ER_tri, ER_topoWet, WC_bio12, ER_annualPET, ER_aridit//ornthwa
## min values :      0,       0,       0,      0,          0,        0,            0,
## max values :      0,       0,       0,      0,          0,        0,            0,
```

Split observations into training and testing

```
pts_split <- rsample::initial_split(
  pts, prop = 0.8, strata = "present")
pts_train <- rsample::training(pts_split)
pts_test <- rsample::testing(pts_split)

pts_train_p <- pts_train %>%
  filter(present == 1) %>%
  as_Spatial()
pts_train_a <- pts_train %>%
  filter(present == 0) %>%
  as_Spatial()
```

Calibrate: Model Selection

```
# calculate variance inflation factor per predictor, a metric of multicollinearity between variables
vif(env_stack)
```

```
##                       Variables       VIF
## 1                        WC_alt  3.797257
## 2                      WC_bio1 24.974187
## 3                      WC_bio2 10.380058
## 4                       ER_tri  4.216269
## 5                    ER_topoWet  4.204632
## 6                      WC_bio12  4.184227
## 7                  ER_annualPET 40.465309
## 8 ER_aridityIndexThornthwaite  3.419468
```

```
# stepwise reduce predictors based on a max correlation of 0.7 (max 1)
v <- vifcor(env_stack, th=0.7)
v
```

```
## 2 variables from the 8 input variables have collinearity problem:
##
## ER_annualPET ER_topoWet
##
## After excluding the collinear variables, the linear correlation coefficients ranges between:
## min correlation ( WC_bio1 ~ WC_alt ):  -0.00675606
## max correlation ( ER_aridityIndexThornthwaite ~ WC_bio12 ):  -0.6604641
##
## ---------- VIFs of the remained variables --------
##                       Variables      VIF
## 1                        WC_alt 2.930131
## 2                      WC_bio1 3.030624
## 3                      WC_bio2 3.747380
## 4                       ER_tri 2.041823
## 5                     WC_bio12 3.922127
## 6 ER_aridityIndexThornthwaite 3.128987
```

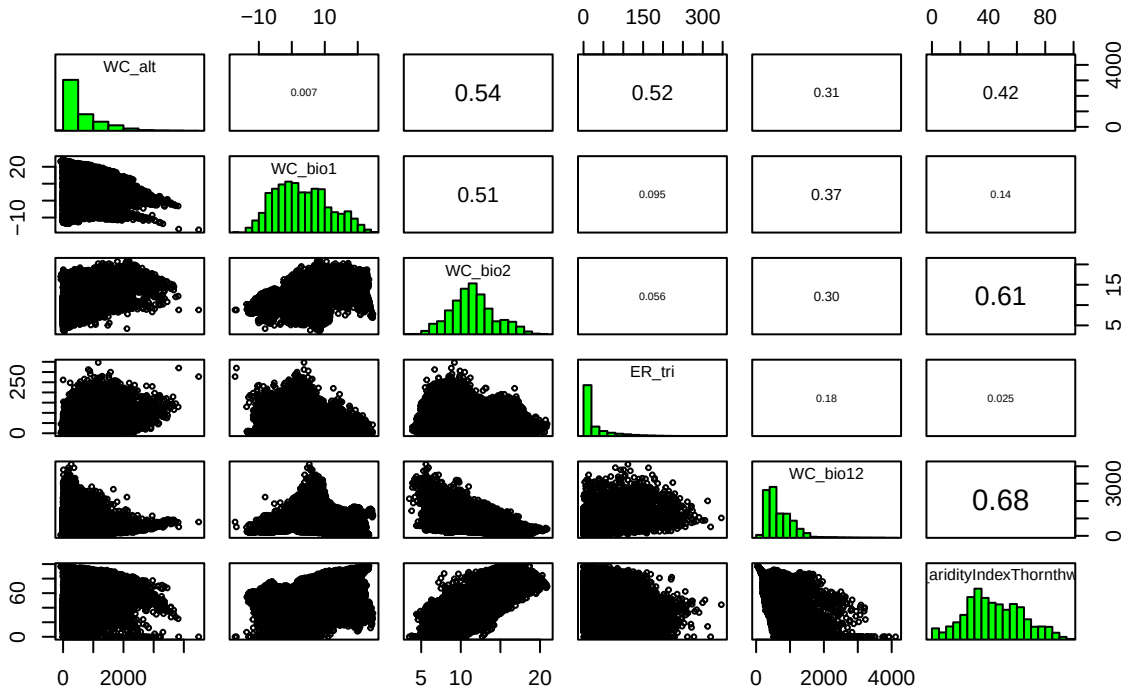
2 variables from the 8 input variables (ER_annualPET and ER_tri) have collinearity problem and were excluded.

After excluding the collinear variables, the linear correlation coefficients ranges between:

min correlation (WC_bio1 ~ WC_alt): -0.001771876

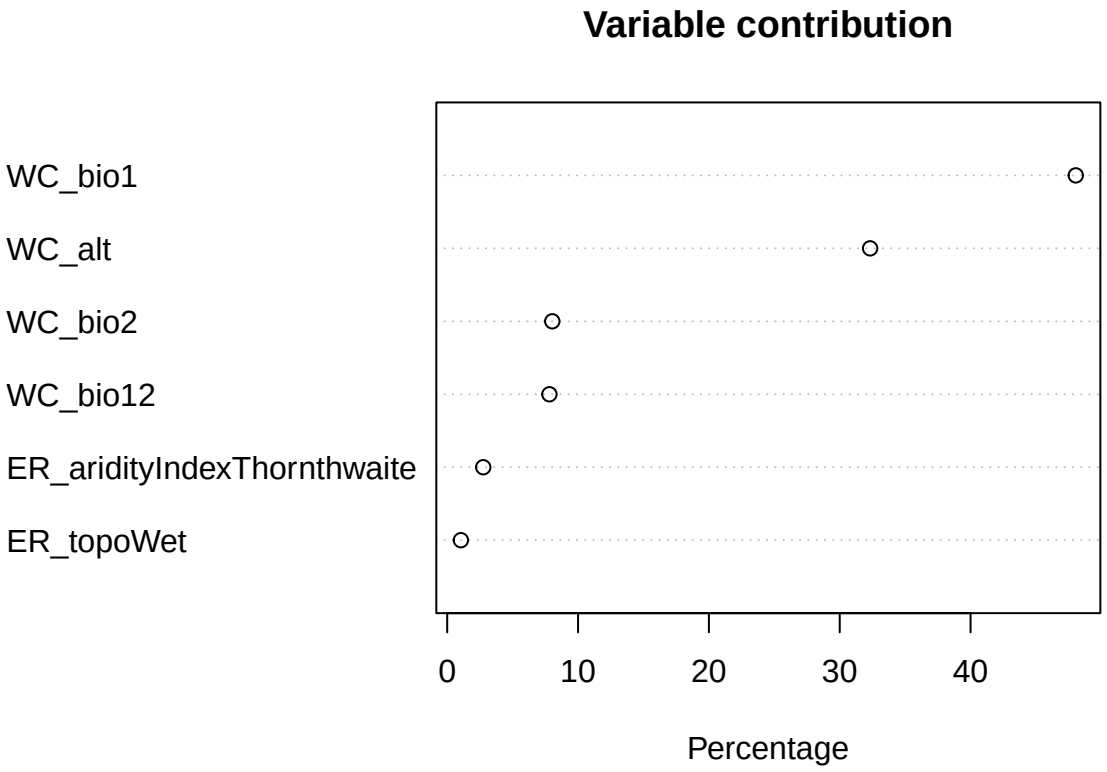
max correlation (ER_aridityIndexThornthwaite ~ WC_bio12): -0.6641443

```
# reduce environmental raster stack to remove collinearity problems
env_stack_v <- usdm::exclude(env_stack, v)

# show pairs plot after multicollinearity reduction with vifcor()
pairs(env_stack_v)
```



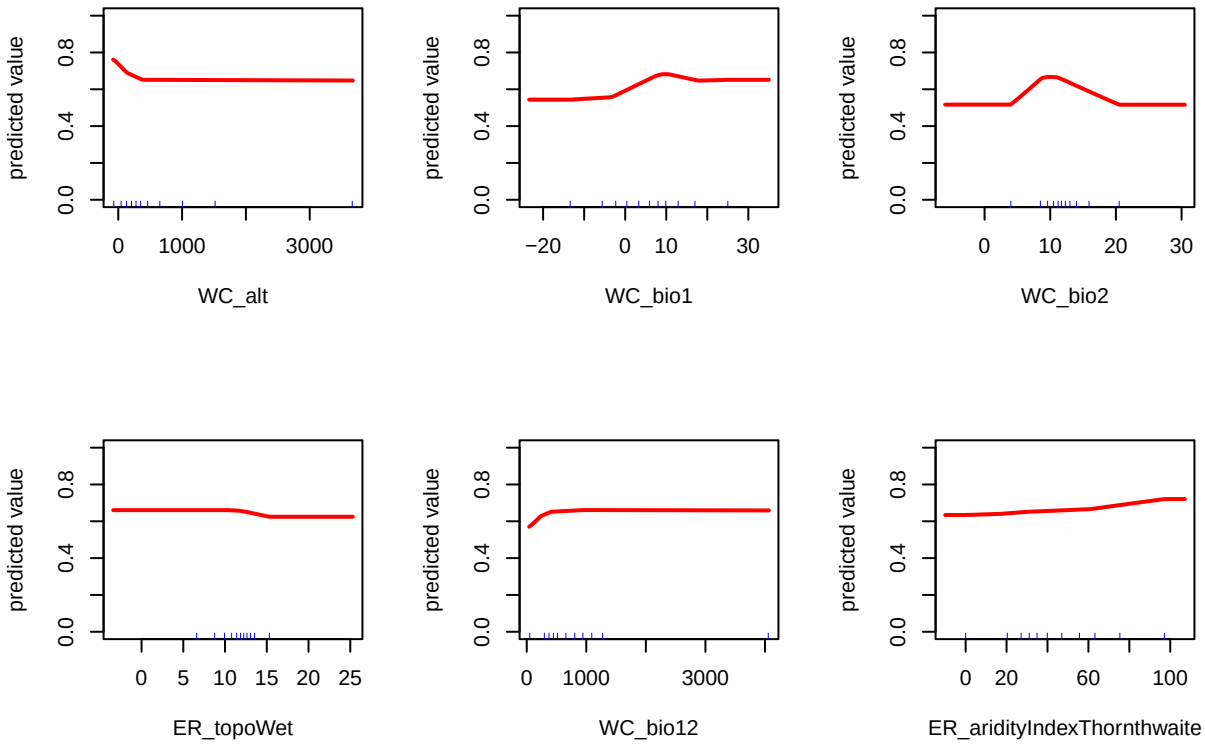

```
# fit a maximum entropy model
if(!file.exists(mdl_maxv_rds)){
  mdl_maxv <- maxent(env_stack_v, sf::as_Spatial(pts_train))
  readr::write_rds(mdl_maxv, mdl_maxv_rds)
}
mdl_maxv <- read_rds(mdl_maxv_rds)
```

```
# plot variable contributions per predictor
plot(mdl_maxv)
```



The most important remaining variable is annual mean temperature (WC_bio1) followed by altitude (WC_alt), mean diurnal temperature range (WC_bio2), then topographic wetness (ER_topoWet).

```
# plot term plots
response(mdl_maxv)
```



Evaluate: Model Performance

Area Under the Curve (AUC), Reciever Operater Characteristic (ROC) Curve and Confustion Matrix

```
pts_test_p <- pts_test %>%
  filter(present == 1) %>%
  as_Spatial()
pts_test_a <- pts_test %>%
  filter(present == 0) %>%
  as_Spatial()

y_maxv <- predict(mdl_maxv, env_stack)


e <- dismo::evaluate(
  p = pts_test_p,
  a = pts_test_a,
  model = mdl_maxv,
  x = env_stack)
e
```

```
## class          : ModelEvaluation
## n presences    : 1996
## n absences     : 1998
## AUC            : 0.886623
## cor            : 0.6649547
## max TPR+TNR at : 0.6503809
```

```
thr <- threshold(e)[['spec_sens']]
thr
```

```
## [1] 0.6503809
```

```
p_true <- na.omit(raster::extract(y_maxv, pts_test_p) >= thr)
a_true <- na.omit(raster::extract(y_maxv, pts_test_a) < thr)
```

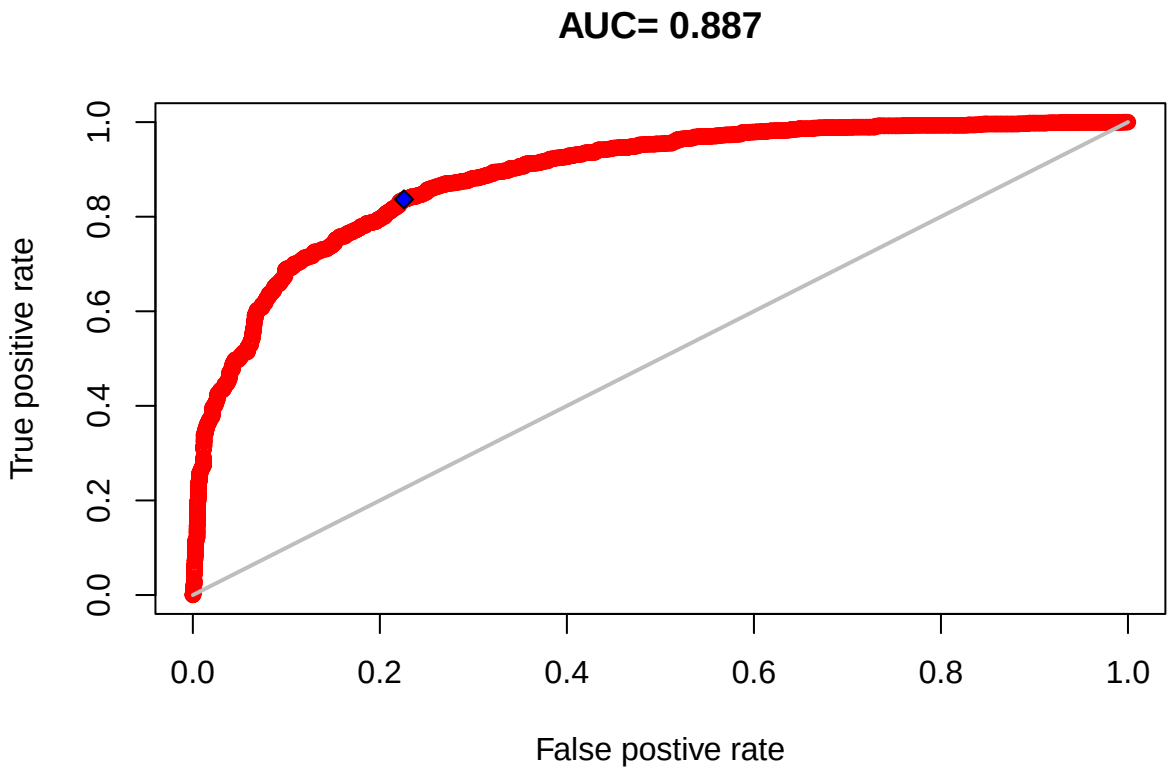
```
# true/false positive/negative rates
tpr <- sum(p_true)/length(p_true)
fnr <- sum(!p_true)/length(p_true)
fpr <- sum(!a_true)/length(a_true)
tnr <- sum(a_true)/length(a_true)
```

```
matrix(
  c(tpr, fnr, fpr, tnr),
  nrow=2, dimnames = list(
    c('present_obs', 'absent_obs'),
    c('present_pred', 'absent_pred')))
```

```
##             present_pred absent_pred
## present_obs    0.8366733   0.2257257
## absent_obs     0.1633267   0.7742743
```

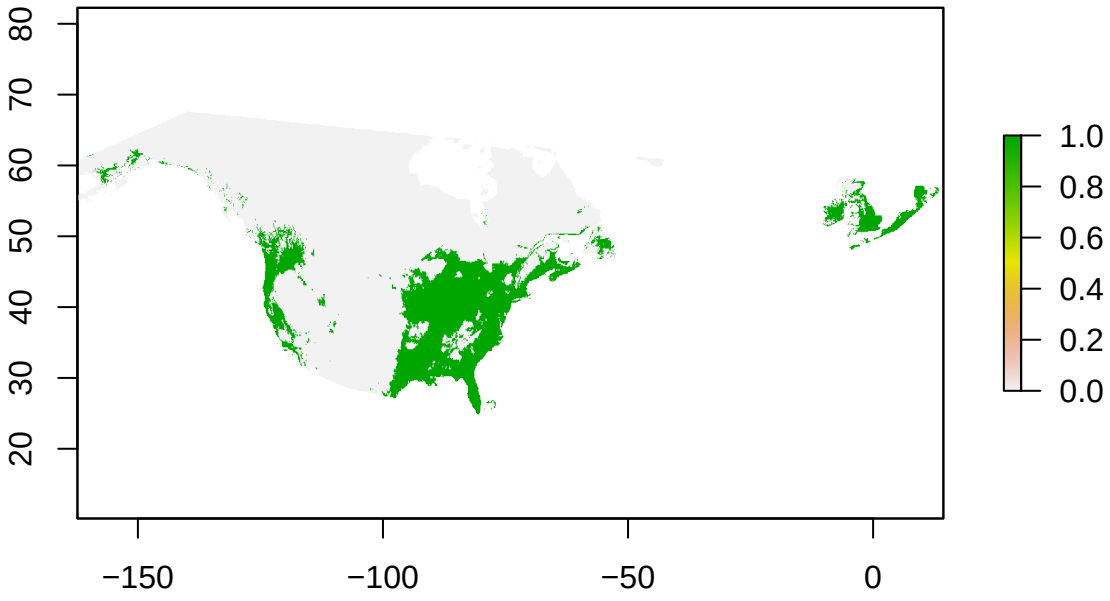
The true positive rate = 83.67%
The true negative rate = 77.43%
The false positive rate = 22.57%
The false negative rate = 16.33%

```
# add point to ROC plot
plot(e, 'ROC')
points(fpr, tpr, pch=23, bg='blue')
```



The Receiver Operator Characteristic (ROC) graph plots the specificity (false positive rate) vs. the sensitivity (true positive rate). Here the diagonal line represents a model that is no better than random guessing. The objective is to maximize the Area Under the Curve (AUC).

```
plot(y_maxv > thr)
```


80
70
60
50
40
30
20
−150
−100
−50
0
1.0
0.8
0.6
0.4
0.2
0.0