Creating Historical and Future Species Distribution Models for Tree-of-Heaven in New England Using R

Andy Atallah

2025-05-21

Introduction

Background of project

Species distribution models (SDMs) aim to show the predicted presence of a species across the landscape (Miller 2010). SDMs are relevant in ecology for studies of how climate change may alter the current ranges of a variety of species (Mainali et al. 2015; Srivistava et al. 2019). For example, a study by Tang et al. (2012) predicts the future forest composition of New England under different modeled climate change scenarios, which has implications for the future of this large region. SDMs are also notably used to model the future spread of invasive plants (Ahmed et al. 2020), which are non-native species that are harmful to native species due to traits such as a fast growth rate (e.g., Graebner et al. 2012).

The importance of removing invasive plants is represented in a U.S. government initiative known as Early Detection and Rapid Response (EDRR; Reaser et al. 2020a). With EDRR, government personnel communicate that invasive species ought to be identified as soon as possible and removed swiftly in order to avoid their deleterious effects on native species (NISIC n.d.). In addition to the threat they pose to native plants, invasive plants are costly to remove (Larson et al. 2011), and land managers may therefore be interested in which species may invade their land in the future, names of which are collected on documents known as watch lists (Jarnevich et al. 2023).

MaxEnt, or maximum entropy modeling, is a type of algorithm used to produce SDMs based only upon points where the species have been observed in the past and explanatory rasters containing values for environmental variables (Phillips et al. 2006). This is the type of model which will be produced in R to predict the distribution of a study species. Point data is readily available from the Global Biodiversity Information Facility (GBIF), a global database of species occurrences, and raster data is available from Worldclim (Fick and Hijmans 2017). Specifically, raster data for bioclimatic variables (i.e., those related to temperature and precipitation) is provisioned in the form of 19 TIFF files each for a historical period (1970-2000) and a series of 20-year periods (2021-2040, 2041-2060, 2061-2080, and 2081-2100).

Importance of project

Invasive species distribution models predict the presence/absence of an invasive species across a study area (e.g., Diao and Wang 2014). The study species for this project is *Ailanthus altissima*, otherwise known as tree-of-heaven. In Massachusetts, this species is one of only 36 plants considered to be invasive (MIPAG 2022), and it can thus be considered a species of relative concern. *A. altissima* is described as especially troublesome among invasive species due to the traits of high seed production, easy distribution of seeds via wind, potential for vegetative growth, and its use of allelopathic effects against native species (Soler and Izquierdo 2024). According to GBIF, tree-of-heaven has not been observed (i.e., there are no presence points) across much of Vermont and New Hampshire and essentially all of Maine (GBIF 2025). As such, one may wonder whether this harmful species could potentially spread into northern New England (the aforementioned three states) from its already-established populations in southern New England (MA, CT, RI), as climate changes.

Project outcomes

The outcome of this project is a script which can be used to remotely download data from examples of two of three basic sources needed for SDM construction: GBIF (presence points) and Worldclim (explanatory rasters). Please see the separate RMarkdown file for documentation of a discrepancy which prohibited the use of a remotely downloaded vector dataset from

ArcGIS Online to delineate the study area. Such an achievement may be possible given connection with ArcGIS Pro, but part of the aim of this project was to create SDMs using open-source or free software.

In keeping with other SDMs created using MaxEnt, the results of the project are a set of rasters showing the predicted probability of presence for tree-of-heaven in different time periods (Liu 2022). There are three rasters shown in the results section: a historical SDM, showing a prediction surface using 1970-2000 data; a future raster, showing a prediction surface using predicted climate data for 2021-2040; and another future raster created from a prediction on 2041-2060 data.

Methods

Point data download

To download point data from GBIF directly to the R session, one can first configure R (https://docs.ropensci.org/rgbif/articles/gbif_credentials.html) to store a GBIF username and password.

```
#library(usethis)
#usethis::edit_r_environ()
```

With the above code, functions from the package `rgbif` will use an existing username and password for GBIF. Downloading data on the website itself always requires authentication, so this is a normal aspect of its provision of species occurrence data. The goal with the `rgbif` package is to attempt to download all presence points of tree-of-heaven within the chosen study region of New England. To do this, one can use the `occ_download` (occurrence download) function in the package to construct a query by supplying various filters to the internal `pred` (predicate) function. When one navigates to the GBIF website to download data, the graphical user interface on the side of the screen adds filters to the query one at a time. This is visible in the URL. For example, when one selects that the occurrence status should be “present,” the term `&occurrence_status=present` is added to the URL. The `occ_download` function works in the same way; users add predicates to the download in order to create a sufficiently selective set of points.

A few filters are included together in the value `pred_default()`, which will be used for simplicity. The added filters are that the occurrence status is present, the points have coordinates, there are no geospatial issues, and the points are not fossil or living specimens (Waller 2020). In addition, codes are supplied to format a query which only extracts points from the study region of New England. Since this area is not a country nor an administrative area, the two built-in options for area filters on GBIF, codes will be used for the six New England states. On GBIF, this looks like:

`gadm_gid=USA.7_1&gadm_gid=USA.46_1&gadm_gid=USA.22_1&gadm_gid=USA.40_1&gadm_gid=USA.30_1&gadm_gid=USA.20_1`.

Disclaimer: The below code chunk was used to download GBIF data on one of the final runs of the script. It successfully downloaded a ZIP file to the data folder for the project once it had run. A subsequent test run, however, suggested a severe lag or rate limit on the GBIF API (the status of the download remained “preparing” for upwards of one hour, when it is supposed to move to “running” within a minute). As such, the below code has been commented out and the ZIP file, which contains a CSV file, will be read into the script to replace the object which would have been remotely downloaded in the final run of the script (during the knitting process).

```
library(rgbif)
library(tidyverse)
```

```
# gbif_download <- occ_download(pred_default(),
#                                pred("taxonKey", 3190653),
#                                pred_or(
#                                  pred("gadm","USA.7_1"),
#                                  pred("gadm","USA.46_1"),
#                                  pred("gadm","USA.22_1"),
#                                  pred("gadm","USA.40_1"),
#                                  pred("gadm","USA.30_1"),
#                                  pred("gadm","USA.20_1")
#                                  ),
#                                format="SIMPLE_CSV"
#                                )
#
# occ_download_wait(gbif_download)
#
# gbif_data <- occ_download_get(gbif_download) %>%
#   occ_download_import()
```

The download was successful, returning 2398 presence points for tree-of-heaven in New England. GBIF data have the potential to change frequently via the addition of points from community science website iNaturalist, whose “research-grade”, or community-vetted, observations make up a significant portion of the records for many species. A benefit of using `rgbif` is that an SDM may be more up-to-date than one which relies on the user manually downloading data at the start of the project.

Here, the download from a past run of the script is loaded into the project due to the aforementioned problem with the GBIF API taking much longer than could be expected.

```
library(readr)
library(sf)
library(here)
```

```
gbif_data <- read_tsv(here("0012198-250515123054153.csv")) # read_csv did not work; https://discourse.gbif.org/t/reading-gbif-downloads-simple-csv-with-r/2615;
```

```
## Rows: 2398 Columns: 50
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: "\t"
## chr  (28): datasetKey, occurrenceID, kingdom, phylum, class, order, family, ...
## dbl  (13): gbifID, individualCount, decimalLatitude, decimalLongitude, coord...
## lgl   (6): infraspecificEpithet, coordinatePrecision, depth, depthAccuracy, ...
## dttm  (3): eventDate, dateIdentified, lastInterpreted
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# https://discourse.gbif.org/t/problem-parsing-large-occurrence-downloads/2570/3

#gbif_data <- st_as_sf(csv_data, coords=c("decimalLongitude", "decimalLatitude"), crs=4326) # Not pursued due to concatenation of coordinates
```

To create an SDM, one is only required to use the latitude and longitude for these points. This is because the final product of an SDM is a prediction surface, or raster (Phillips et al. 2006) – thus, the characteristics of each point, such as the date associated with each observation on GBIF, are not considered important.

```
# Select only lon/lat columns
gbif_data_clean <- gbif_data[,c("decimalLongitude", "decimalLatitude")]
```

Raster download

Historical bioclimatic raster

Raster data from Worldclim has the benefit of being relatively more static. The Worldclim raster for historical bioclimatic variables is perhaps not likely to change frequently because it is a product which has already been developed. However, there are still methods to obtain the data without manually downloading the files. One major benefit of downloading historical Worldclim data using the `geodata` package (https://rdrr.io/cran/geodata/man/worldclim.html) is that one is able to specify a 30 degree tile instead of downloading data for the entire globe. This should lessen the amount of time needed to download the files. Though data can be downloaded at a very high resolution of 30 seconds, this could be a prohibitively high cost in terms of file storage on some hardware. Here, the `res` argument is instead included as `2.5` minutes in order to lessen space constraints and computational time. Note that the `worldclim_tile` function only seemed to download the 30 second resolution raster, so the global raster at the lower resolution of 2.5 minutes had to be downloaded instead using `worldclim_global`.

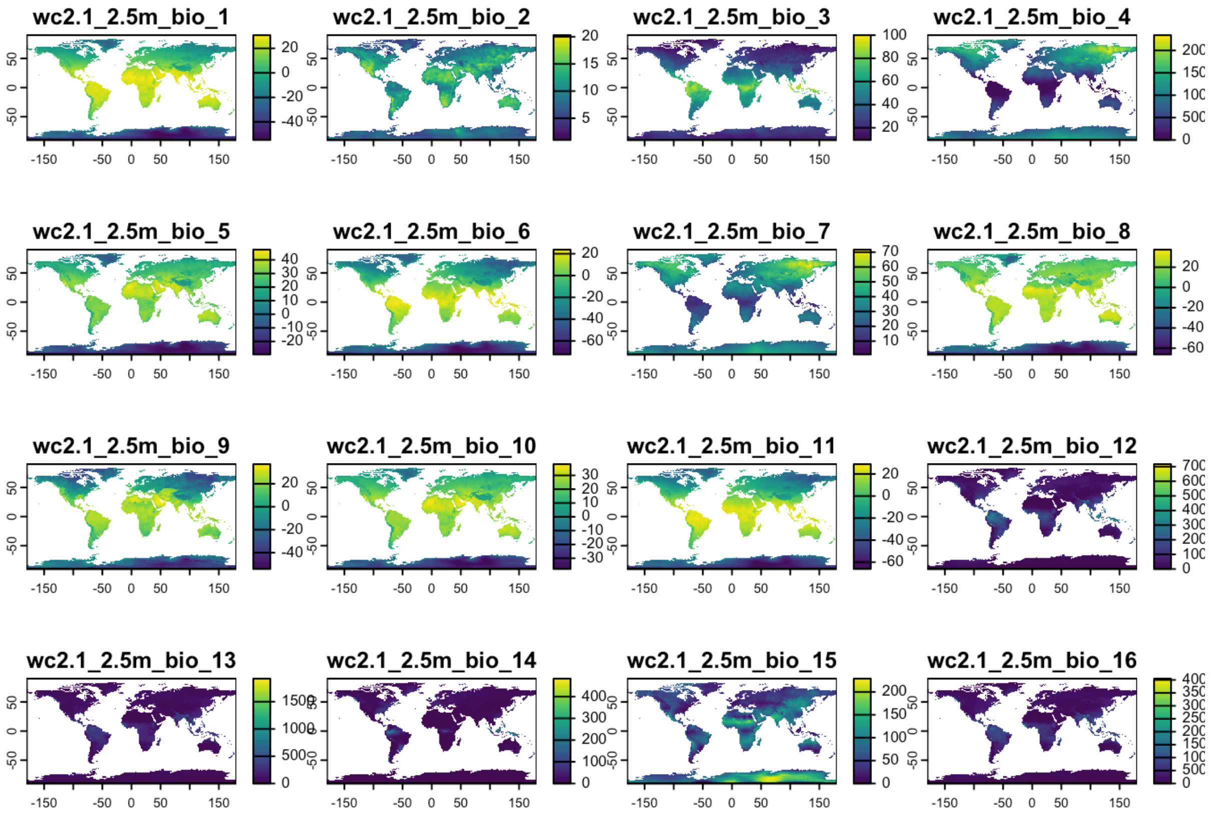
```r
library(geodata)
library(terra)
```

```r
# Set resolution as local variable
res=2.5
lon=-72
lat=42

# Configure download of historical (i.e., 1970-2000) bioclimatic data
bc_rast_historical_25 <- worldclim_global(
                                     var="bio",
                                     res=res,
                                     path="data",
                                     version="2.1"
                                     )
```

After viewing the below plot, it is clear that the above function successfully downloaded the bioclimatic raster data – all 19 bands. The raster stack can be cropped later.

```r
# Test out to see if this downloaded correctly
terra::plot(bc_rast_historical_25)
```



Future bioclimatic rasters

There is a different function in the `geodata` package which is used to download future climate data from Worldclim – `cimp6_world` (or `cimp6_tile`). The documentation for the function (https://rdrr.io/github/rspatial/geodata/man/cmip6.html) shows that arguments passed to the function are in line with the many different options available for manually downloading future data on the Worldclim website. In addition to resolution, which was covered above, there is also a `model` argument. This corresponds to various GCMs, or global climate models, created by different organizations. The model GFDL-ESM4 will be used in this script because it was created by NOAA. Next, there is an `ssp` argument. This corresponds to a Shared Socio-economic Pathway (SSP) which projects greenhouse gas emissions under different global scenarios (Meinshausen et al. 2020). The available options are `126`, `245`, `370`, and `585`. The option `370` will be chosen because it is in the middle of the options and is thus not the most optimistic nor pessimistic scenario for the future. Last, there is a `time` argument which corresponds to the future time period covered by the data. There are options for 2021-2040, 2041-2060, or 2061-2080. The first two options will be chosen here to represent periods perhaps most relevant for short- and mid-term planning.

Testing differences in SDMs created from all combinations of these options would be intriguing, but it would require a large amount of storage and computational time due to the large file sizes, especially if using the highest resolution available is desired. It does not seem possible to restrict the future downloads to one country using the `geodata` package. The `cimp6_tile` function is tested below, but it returns an error.

```
# Set local variables
#lon=-72
#lat=42
model="GFDL-ESM4"
ssp="370"
time_1="2021-2040"
time_2="2041-2060"
res=2.5

# Download 2021-2040 data (tile)
bc_rast_2021_2040 <- cmip6_tile(lon=lon,
                                lat=lat,
                                model=model,
                                ssp=ssp,
                                time=time_1,
                                var="bioc",
                                res=res,
                                path="data")
```

```
## The geodata server is temporary out of service for maintenance. It should be back on 20 June.
```

```
## download failed
```

```
# Download 2041-2060 data (tile)
bc_rast_2041_2060 <- cmip6_tile(lon=lon,
                                lat=lat,
                                model=model,
                                ssp=ssp,
                                time=time_2,
                                var="bioc",
                                res=res,
                                path="data")
```

```
## The geodata server is temporary out of service for maintenance. It should be back on 20 June.
## download failed
```

As seen above, the `cimp6_tile` function does not appear to work at the moment. Maintenance appears to be a reoccurring issue (https://github.com/rspatial/geodata/issues/61). The global data will instead be downloaded.

```
# Set local variables
model="GFDL-ESM4"
ssp="370"
time_1="2021-2040"
time_2="2041-2060"
res=2.5

# Download 2021-2040 data (world)
bc_rast_2021_2040_world <- cmip6_world(
                                model=model,
                                ssp=ssp,
                                time=time_1,
                                var="bioc",
                                res=res,
                                path="data"
                                )

# Download 2041-2060 data (world)
bc_rast_2041_2060_world <- cmip6_world(
                                model=model,
                                ssp=ssp,
                                time=time_2,
                                var="bioc",
                                res=res,
                                path="data"
                                )
```

Data preprocessing

Raster masking and cropping

It is important that these three raster stacks are cropped to the study area before other steps are pursued to reduce computational time.

Now, a study area shapefile or vector dataset is needed to mask/crop these raster stacks. The chosen study area is New England. Attempting to use the `arcgis` package to call down a layer from ArcGIS Online was promising but did not function appropriately, perhaps due to the requirement (https://developers.arcgis.com/r-bridge/installation/) to have a license of ArcGIS Pro with which to connect R and install the `arcgisbinding` package. Said package has a function called `arc.data2sf` (https://www.rdocumentation.org/packages/arcgisbinding/versions/1.0.1.229/topics/arc.data2sf) which may be necessary for a feature layer from ArcGIS Online to be properly read as an `sf` object. The normal workflow of downloading a shapefile and reading it in with the `sf` package can instead be pursued. The authoritative source used for this purpose was the U.S. Census Bureau Cartographic Boundary Files (https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html) website.

```
library(sf)
library(tidyverse)
```

```
# Read in Census Divisions shapefile
census_divisions = st_read(here::here('data',
                                  'cb_2018_us_division_500k',
                                  'cb_2018_us_division_500k.shp'))
```

```
## Reading layer `cb_2018_us_division_500k' from data source 
##   `/Users/a/Desktop/Other/SDAR/Final2/data/cb_2018_us_division_500k/cb_2018_us_division_500k.sh
p' 
##   using driver `ESRI Shapefile'
## Simple feature collection with 9 features and 7 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -179.1489 ymin: 18.91036 xmax: 179.7785 ymax: 71.36516
## Geodetic CRS:  NAD83
```

```
new_england <- census_divisions %>%
  filter(NAME == "New England")
```

Reprojection of the shapefile into the CRS of the `SpatRaster` would be beneficial.

```
# Reproject the shapefile
new_england_proj <- st_transform(new_england, crs(bc_rast_historical_25))
new_england_proj
```

	DIVISIONCE	AFFGEOID	GE...	NAME	LS...	ALAND	AWATER	geometry
	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<s_MULTIP>
1	1	0300000US1	1	New England	69	162376417481	24072855206	<s_MULTIP>

1 row

```
# Mask, crop, and plot historical raster to test workflow
bc_historical_masked <- mask(bc_rast_historical_25, new_england_proj)
bc_historical_clip <- crop(bc_historical_masked, new_england_proj)
#plot(bc_historical_clip[[1]])
```

It seems as though the above workflow functions appropriately, so it will be repeated with the two future `SpatRasters` .

```
# Mask, crop, and plot future rasters

# 2021-2040
bc_2021_2040_masked <- mask(bc_rast_2021_2040_world, new_england_proj)
bc_2021_2040_clip <- crop(bc_2021_2040_masked, new_england_proj)

# 2041-2060
bc_2041_2060_masked <- mask(bc_rast_2041_2060_world, new_england_proj)
bc_2041_2060_clip <- crop(bc_2041_2060_masked, new_england_proj)
```

To run the next portion of the workflow, the `ENMevaluate` function from the package `ENMeval` , it seems as though a `RasterStack` object is needed. This can be achieved by converting the `terra` object known as a `SpatRaster` to the `raster` object known as a `RasterStack` .

```
# Convert to RasterStack
bc_historical_stack <- raster::stack(bc_historical_clip)
bc_historical_stack
```

MaxEnt tuning

Now, the `ENMevaluate` function can be run. The purpose of this function can be understood as part of the concept of machine learning. For certain computer algorithms, users are intended to supervise performance in order to determine the best combination of parameters for the model to predict on new data. An example is a random forest algorithm which is intended to predict the presence or absence of a feature at various locations. The following function, `ENMevaluate`, tests parameter sets for the `maxnet` algorithm, which is the R implementation of `MaxEnt`. In theory, this saves a great amount of time which the user might otherwise spend inputting different parameter combinations into the `dismo::maxent` function itself. The feature combination (FC) and regularization multipler (RM) which cause the model to perform best will be extracted from a dataframe produced by the `ENMevaluate` function.

Certain feature combinations are commented out due to warning and/or error messages produced when all, including just hinge (H), were included. This may be due to clustering in the GBIF points which place many points in the same raster cells.

```
library(ENMeval)
```

```
  enmeval_results_5deg = ENMevaluate(occ = gbif_data_clean,
                                    env = bc_historical_stack,
                                    bg = NULL,
                                    tune.args = list(fc = c("L",
                                                            "LQ",
                                                            #"H",
                                                            "LQH"
                                                            #"LQHP",
                                                            #"LQHPT"
                                                            ),
                                                 rm = 1:3),
                                   partitions = "randomkfold",
                                   partition.settings = list(kfolds = 5),
                                   algorithm = "maxnet",
                                   taxon.name = "Ailanthus altissima")
```

```
# Extract results dataframe
enmeval_df = enmeval_results_5deg@results

# View ordered by delta.AICc, which reveals model performance
enmeval_df %>%
  arrange(delta.AICc)
```

fc	rm	tune.args	auc.train	cbi.train	auc.diff.avg	auc.diff.sd	auc.val.avg	auc.val.sd
<fct>	<fct>	<fct>	<dbl>	<lgl>	<dbl>	<dbl>	<dbl>	<dbl>
LQH	2	fc.LQH_rm.2	0.9429996	NA	0.005177543	0.003979179	0.9398797	0.005151510
LQH	1	fc.LQH_rm.1	0.9447814	NA	0.005498704	0.004057822	0.9404037	0.004000152
LQH	3	fc.LQH_rm.3	0.9411553	NA	0.005521488	0.003679094	0.9391925	0.005724096
LQ	1	fc.LQ_rm.1	0.9382778	NA	0.005980980	0.005206068	0.9371535	0.006828324
LQ	2	fc.LQ_rm.2	0.9374100	NA	0.005678222	0.005703373	0.9366062	0.006868309
LQ	3	fc.LQ_rm.3	0.9371692	NA	0.005628925	0.005619931	0.9361884	0.006702218
L	1	fc.L_rm.1	0.9369837	NA	0.006086108	0.006109791	0.9363183	0.007323300
L	2	fc.L_rm.2	0.9369145	NA	0.006039770	0.006086079	0.9362805	0.007277062
L	3	fc.L_rm.3	0.9367933	NA	0.006081985	0.005889220	0.9361793	0.007217526

9 rows | 1-9 of 19 columns

It appears that the best model has the tuning arguments LQH (linear, quadratic, and hinge) and regularization multiplier 2. This model has a training area under the ROC curve (AUC) value of 0.94 (in the first run of this function). The appropriate parameters can be extracted from the data frame to prepare for running the `dismo::maxent` function.

```
# Subset the ENMeval results to get the best model
enm_bestmodel <- subset(enmeval_df, delta.AICc == 0)

# Extract the FC and RM parameters
maxent_fc <- as.character(enm_bestmodel$fc)
maxent_rm <- as.character(enm_bestmodel$rm)
```

MaxEnt

With the MaxEnt model tuned before the `dismo::maxent` function has even been run, time has theoretically been saved in tuning to find the appropriate combination of FC and RM values. Now, a combination of inputs will be used to train a MaxEnt model for subsequent prediction on the historical data.

```
library(rJava)
```

```
# Run MaxEnt model using RasterStack, points, FC, and RM
maxent_model_toh <- dismo::maxent(bc_historical_stack, as.matrix(gbif_data_clean),
                             features = maxent_fc,
                             betamultiplier = maxent_rm)
```

Now, one can use the MaxEnt model to predict the probability of presence using the historical raster stack. This will produce a raster.

```
# Predict tree-of-heaven distribution in New England
sdm_curr <- dismo::predict(maxent_model_toh, bc_historical_stack)
```

After Liu (2022) and Wimberly (2023), the same model will be used with the future raster stacks in order to predict on future data. First, raster stacks must be created in line with how `bc_historical_stack` was created above.

```
# Create raster stacks
bc_2021_2040_stack <- raster::stack(bc_2021_2040_clip)
bc_2041_2060_stack <- raster::stack(bc_2041_2060_clip)
```

Checking the names of the historical stack reveals the pattern `wc2.1_2.5m_bio_n` .

```
hist_names <- names(bc_historical_stack)
str(hist_names)
```

```
##  chr [1:19] "wc2.1_2.5m_bio_1" "wc2.1_2.5m_bio_2" "wc2.1_2.5m_bio_3" ...
```

However, in the future stacks, the names are `wc2.1_2.5m_bioc_GFDL.ESM4_ssp370_2021.2040_n` . This will throw an error when the MaxEnt function is run due to incompatibility of names.

```
names(bc_2021_2040_stack)
```

```
##  [1] "wc2.1_2.5m_bioc_GFDL.ESM4_ssp370_2021.2040_1"
##  [2] "wc2.1_2.5m_bioc_GFDL.ESM4_ssp370_2021.2040_2"
##  [3] "wc2.1_2.5m_bioc_GFDL.ESM4_ssp370_2021.2040_3"
##  [4] "wc2.1_2.5m_bioc_GFDL.ESM4_ssp370_2021.2040_4"
##  [5] "wc2.1_2.5m_bioc_GFDL.ESM4_ssp370_2021.2040_5"
##  [6] "wc2.1_2.5m_bioc_GFDL.ESM4_ssp370_2021.2040_6"
##  [7] "wc2.1_2.5m_bioc_GFDL.ESM4_ssp370_2021.2040_7"
##  [8] "wc2.1_2.5m_bioc_GFDL.ESM4_ssp370_2021.2040_8"
##  [9] "wc2.1_2.5m_bioc_GFDL.ESM4_ssp370_2021.2040_9"
## [10] "wc2.1_2.5m_bioc_GFDL.ESM4_ssp370_2021.2040_10"
## [11] "wc2.1_2.5m_bioc_GFDL.ESM4_ssp370_2021.2040_11"
## [12] "wc2.1_2.5m_bioc_GFDL.ESM4_ssp370_2021.2040_12"
## [13] "wc2.1_2.5m_bioc_GFDL.ESM4_ssp370_2021.2040_13"
## [14] "wc2.1_2.5m_bioc_GFDL.ESM4_ssp370_2021.2040_14"
## [15] "wc2.1_2.5m_bioc_GFDL.ESM4_ssp370_2021.2040_15"
## [16] "wc2.1_2.5m_bioc_GFDL.ESM4_ssp370_2021.2040_16"
## [17] "wc2.1_2.5m_bioc_GFDL.ESM4_ssp370_2021.2040_17"
## [18] "wc2.1_2.5m_bioc_GFDL.ESM4_ssp370_2021.2040_18"
## [19] "wc2.1_2.5m_bioc_GFDL.ESM4_ssp370_2021.2040_19"
```

Since the historical names entered the environment as a vector, they can be used to replace the names of the 19 bands of the future rasters.

```
# Change names of future rasters; https://gis.stackexchange.com/questions/452710/writeraster-how-to
-mantain-layer-names-when-writing-raster-data-to-a-file
names(bc_2021_2040_stack) <- hist_names
names(bc_2021_2040_stack)
```

```
##  [1] "wc2.1_2.5m_bio_1"  "wc2.1_2.5m_bio_2"  "wc2.1_2.5m_bio_3"
##  [4] "wc2.1_2.5m_bio_4"  "wc2.1_2.5m_bio_5"  "wc2.1_2.5m_bio_6"
##  [7] "wc2.1_2.5m_bio_7"  "wc2.1_2.5m_bio_8"  "wc2.1_2.5m_bio_9"
## [10] "wc2.1_2.5m_bio_10" "wc2.1_2.5m_bio_11" "wc2.1_2.5m_bio_12"
## [13] "wc2.1_2.5m_bio_13" "wc2.1_2.5m_bio_14" "wc2.1_2.5m_bio_15"
## [16] "wc2.1_2.5m_bio_16" "wc2.1_2.5m_bio_17" "wc2.1_2.5m_bio_18"
## [19] "wc2.1_2.5m_bio_19"
```

```
names(bc_2041_2060_stack) <- hist_names
names(bc_2041_2060_stack)
```

```
##  [1] "wc2.1_2.5m_bio_1"  "wc2.1_2.5m_bio_2"  "wc2.1_2.5m_bio_3"
##  [4] "wc2.1_2.5m_bio_4"  "wc2.1_2.5m_bio_5"  "wc2.1_2.5m_bio_6"
##  [7] "wc2.1_2.5m_bio_7"  "wc2.1_2.5m_bio_8"  "wc2.1_2.5m_bio_9"
## [10] "wc2.1_2.5m_bio_10" "wc2.1_2.5m_bio_11" "wc2.1_2.5m_bio_12"
## [13] "wc2.1_2.5m_bio_13" "wc2.1_2.5m_bio_14" "wc2.1_2.5m_bio_15"
## [16] "wc2.1_2.5m_bio_16" "wc2.1_2.5m_bio_17" "wc2.1_2.5m_bio_18"
## [19] "wc2.1_2.5m_bio_19"
```

Now, MaxEnt can be applied to the future rasters.

```
# Predict tree-of-heaven distribution in New England for the period 2021-2040
sdm_future_1 <- dismo::predict(maxent_model_toh, bc_2021_2040_stack)
```

```
# Predict tree-of-heaven distribution in New England for the period 2041-2060
sdm_future_2 <- dismo::predict(maxent_model_toh, bc_2041_2060_stack)
```

Results

The SDM rasters can be plotted to reveal the differences between them.

```r
library(raster)
```

```r
# Convert raster to a dataframe
sdm_curr_df <- as.data.frame(rasterToPoints(sdm_curr), stringsAsFactors = FALSE)
# Basic plot
  ggplot() +
    geom_tile(data = sdm_curr_df, aes(x = x, y = y, fill = layer)) +
    coord_fixed(ratio = 1) +  # Preserve aspect ratio
    labs(fill = "Value", x = "Longitude", y = "Latitude") +
    scale_fill_viridis_c(name="Probability of Presence\n (1970-2000)")
```

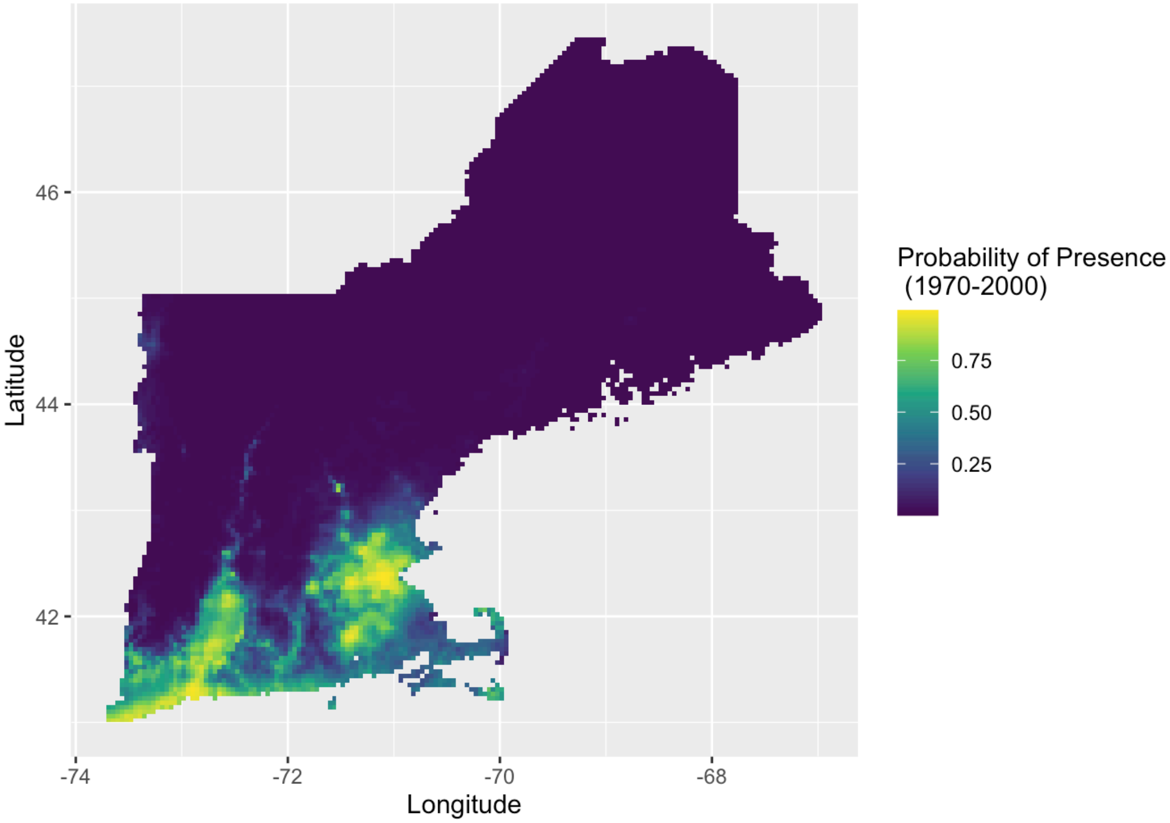


Figure 1. Historical species distribution model (SDM) for the species Ailanthus altissima (tree-of-heaven) in New England. SDM created using MaxEnt using Worldclim historical climate data (1970-2000) for 19 bioclimatic variables.

```r
# Convert raster to a dataframe
sdm_fut1_df <- as.data.frame(rasterToPoints(sdm_future_1), stringsAsFactors = FALSE)
# Basic plot
  ggplot() +
    geom_tile(data = sdm_fut1_df, aes(x = x, y = y, fill = layer)) +
    coord_fixed(ratio = 1) +  # Preserve aspect ratio
    labs(fill = "Value", x = "Longitude", y = "Latitude") +
    scale_fill_viridis_c(name="Probability of Presence\n (2021-2040)")
```

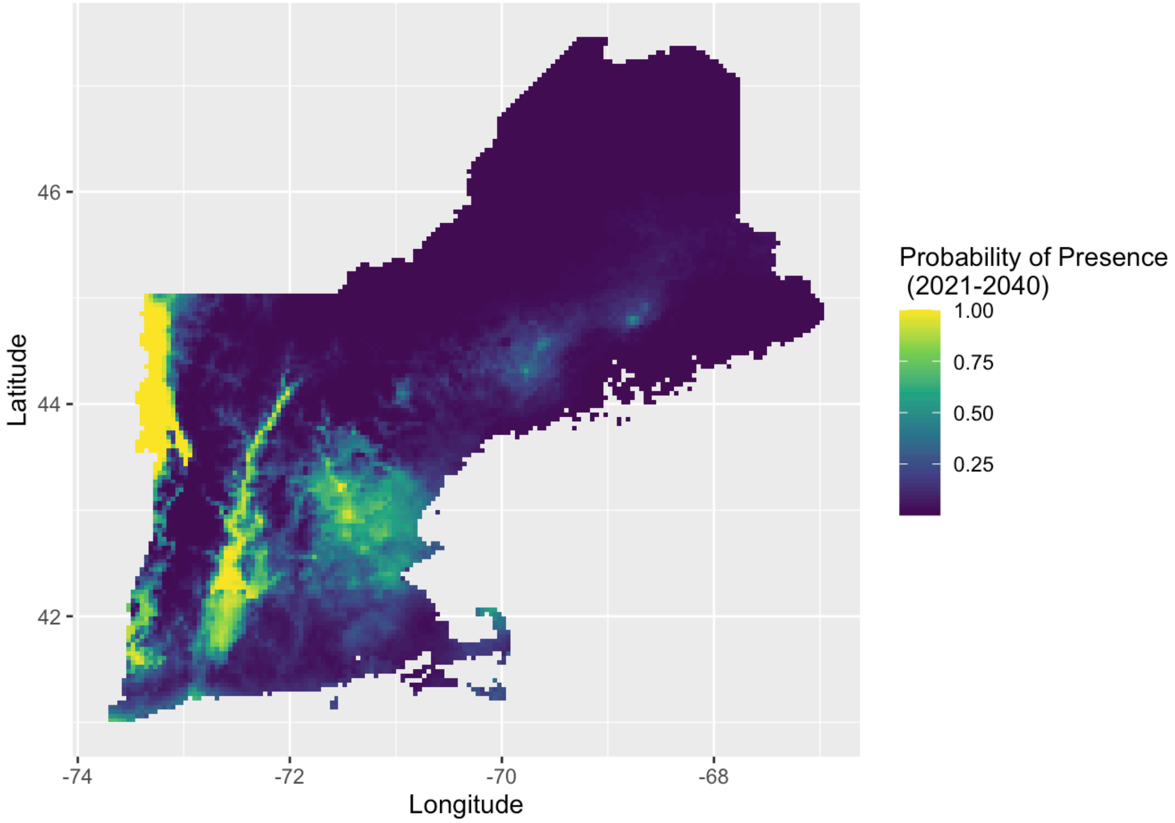


Figure 2. Future species distribution model (SDM) for the species Ailanthus altissima (tree-of-heaven) in New England in the period 2021-2040. SDM created using MaxEnt using Worldclim future climate data (GCM = GFDL-ESM4; SSP = 370) for 19 bioclimatic variables.

```
# Convert raster to a dataframe
sdm_fut2_df <- as.data.frame(rasterToPoints(sdm_future_2), stringsAsFactors = FALSE)
# Basic plot
  ggplot() +
    geom_tile(data = sdm_fut2_df, aes(x = x, y = y, fill = layer)) +
    coord_fixed(ratio = 1) +  # Preserve aspect ratio
    labs(fill = "Value", x = "Longitude", y = "Latitude") +
    scale_fill_viridis_c(name="Probability of Presence\n (2041-2060)")
```

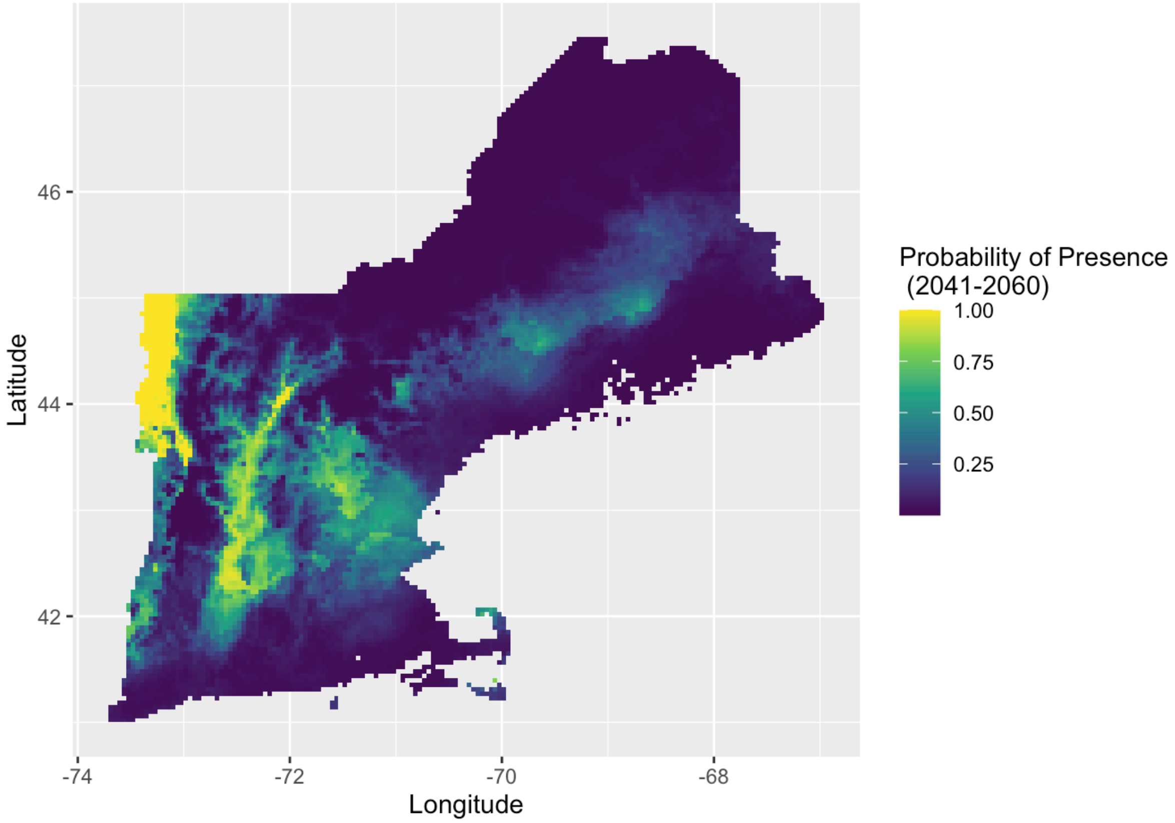


Figure 3. Future species distribution model (SDM) for the species Ailanthus altissima (tree-of-heaven) in New England in the period 2041-2060. SDM created using MaxEnt using Worldclim future climate data (GCM = GFDL-ESM4; SSP = 370) for 19 bioclimatic variables.

The legend titles for the plots were assigned based on Liu (2022).

Using historical climate data, tree-of-heaven is predicted to be found in: eastern Massachusetts, including the Islands; southwestern Connecticut; and Rhode Island. This comprises southern New England. There is a very low predicted probability of presence in northern New England, including almost a 0% predicted probability of presence across all of Maine (Figure 1).

In the nearest future period, 2021-2040, the predicted range of the species changes. In fact, there is somewhat of a northward shift. The model is notably near-certain that tree-of-heaven could be found in western Vermont, particularly near Lake Champlain, and presence is deemed likely (e.g., approximately 75%) in much of eastern New Hampshire and southern Maine. Elsewhere in northern New England, high presence probability appears to follow the Connectcut River, which lies near the border of New Hampshire and Vermont. There are also pockets of middling probability of presence in inland central Maine. In the south, however, there is less confidence in the presence of the species on the South Shore of Massachusetts and southwestern Connecticut (Figure 2).

The prediction surface for the latest future period included in this project, 2041-2060, shows a further northward shift in predicted presence probability. The areas with middling probability of presence (approximately 50-75%) in inland Maine have expanded, and areas with similar values are found in veins in central and western Vermont and central New Hampshire. Meanwhile, the species is no longer predicted to be present in the southernmost area of the raster, corresponding to southern Connecticut, southern Massachusetts, and most of Rhode Island (Figure 3).

Discussion

Historical range

The raster showing predicted historical distribution (Figure 1), which is based on historical data, is theoretically distinct from the GBIF point data because it displays the predicted distribution of the species rather than the observed locations. This may not have the exact same extent as the presence points due to the concept that the species is not found in every place it may be predicted to be found. Indeed, when comparing Figure 1 to Figure 4, displayed below, it is apparent that there are some slight differences. For example, the species appears to be predicted to be present with a probability of 50% or greater in eastern Connecticut, where there have not been many observations of the species to date (Figure 1, Figure 4). This may mean that other factors besides the 19 bioclimatic variables, such as land use (Liu 2022), may have some influence in predicting the species distribution.

```
ggplot() +
  geom_sf(data = new_england) +
  geom_point(data = gbif_data_clean, aes(x=decimalLongitude, y=decimalLatitude), size=0.5, alpha=0.
5) +
  labs(x="Longitude", y="Latitude")
```

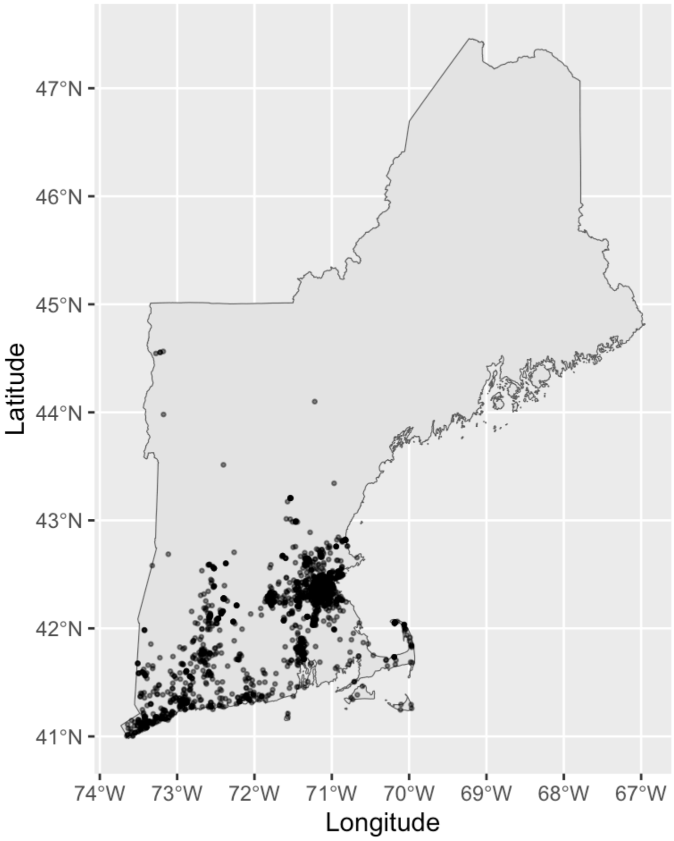


Figure 4. Distribution of GBIF presence points in New England. Download occurred on May 21, 2025.

Historical vs. future range

As noted in the results section, there are large observable differences in the three prediction surfaces (Figure 1-3). Together, these differences suggest that the values of the bioclimatic rasters in the future periods are meaningfully different than those in the historical (1970-2000) raster stack. The model predicts based on the extracted raster values from the cells in which points lie (Phillips et al. 2006). It is thus possible that the bioclimatic raster values in southern New England change such that the model loses confidence in the predicted presence and that the raster values in northern New England change such that they are similar to the values in the cells in which points lie in Figure 1. The latter condition is perhaps simpler to comprehend as a function of a projected increase in annual temperature. In climate simulations reported by Tang et al. (2012), temperature increased in New England, causing deciduous forests to shift northward. Janowiak et al. (2018) also report that climate models suggest significant increases in temperature by 2100 across New England. In the list of layer names (https://www.worldclim.org/data/bioclim.html) for Worldclim bioclimatic data, there are many which relate to temperature; it is split into two between those variables related to temperature and those related to precipitation. One can start by investigating changes in variable 1, annual mean temperature.

```
# Subtract latest raster from historical raster
minus_bio_1 <- bc_2041_2060_stack[[1]] - bc_historical_stack[[1]]

# Convert raster to a dataframe
minus_bio_1_df <- as.data.frame(rasterToPoints(minus_bio_1), stringsAsFactors = FALSE)

# Plot difference raster
ggplot() +
  geom_tile(data = minus_bio_1_df, aes(x = x, y = y, fill = layer)) +
    coord_fixed(ratio = 1) +  # Preserve aspect ratio
    labs(fill = "Value", x = "Longitude", y = "Latitude") +
    scale_fill_viridis_c(name="BIO 1: Annual Mean Temperature \n(°C): 2041-2060 - 1970-2000")
```

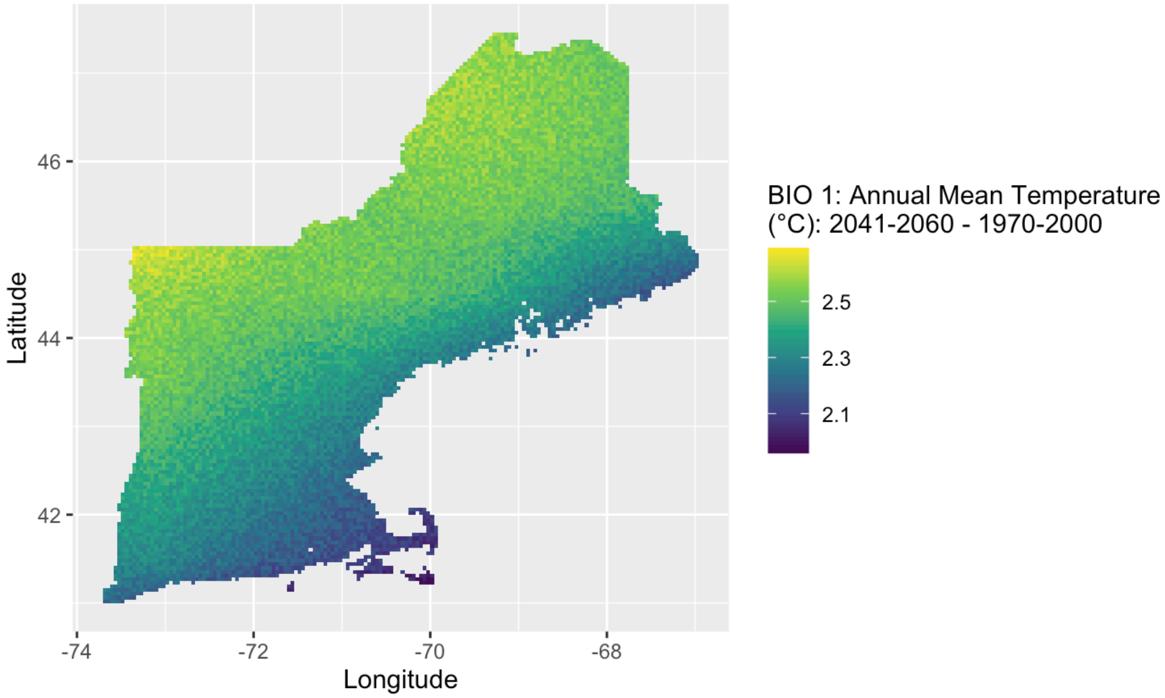


Figure 5: Difference raster created by subtracting band 1 of the historical raster stack (1970-2000 data) from band 1 of the future raster stack for 2041-2060. This shows Worldclim bioclimatic variable 1, or annual mean temperature.

As can be seen in Figure 5, annual mean temperature increases most in northwestern Vermont, which could be a reason for the high predicted presence probability in this region in Figures 2-3. It also increases least in the southern CT, MA, and RI, which may be relevant to the lower presence probability predicted for this area in Figures 2-3.

The variable of annual precipitation (BIO 12) can be tested as a second variable to compare with BIO 1.

```r
# Subtract latest raster from historical raster
minus_bio_12 <- bc_2041_2060_stack[[12]] - bc_historical_stack[[12]]

# Convert raster to a dataframe
minus_bio_12_df <- as.data.frame(rasterToPoints(minus_bio_12), stringsAsFactors = FALSE)

# Plot difference raster
ggplot() +
  geom_tile(data = minus_bio_12_df, aes(x = x, y = y, fill = layer)) +
    coord_fixed(ratio = 1) +  # Preserve aspect ratio
    labs(fill = "Value", x = "Longitude", y = "Latitude") +
    scale_fill_viridis_c(name="BIO 12: Annual Precipitation \n(mm): 2041-2060 - 1970-2000")
```

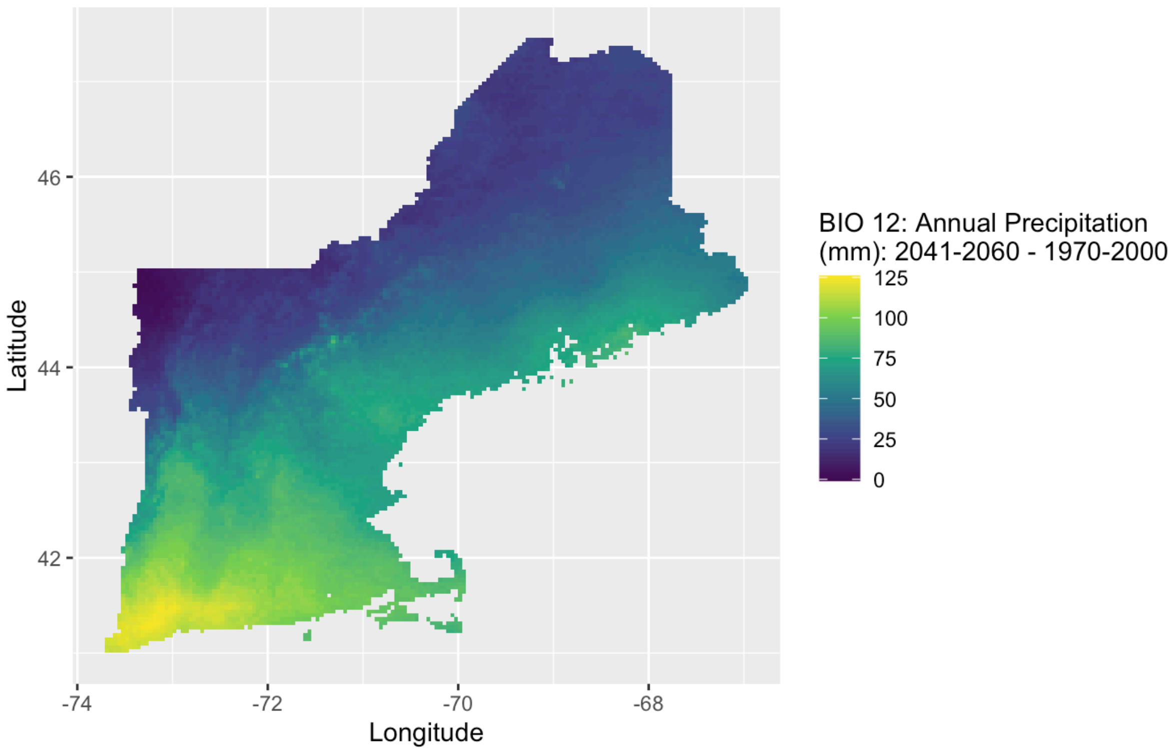


Figure 6: Difference raster created by subtracting band 12 of the historical raster stack (1970-2000 data) from band 12 of the future raster stack for 2041-2060. This shows Worldclim bioclimatic variable 12, or annual precipitation.

Projected increases in annual precipitation (BIO 12) appear highest in southwestern Connecticut and lowest in northwestern Vermont (Figure 6). This observation is notable in the context of Figure 2-3, which show an increased presence probability and decreased presence probability, respectively, for these same regions compared to Figure 1. The annual precipitation values increasing in southern New England could be one factor of why tree-of-heaven is not predicted to be found there by the MaxEnt model in future periods, especially 2041-2060 (Figure 3).

It would be beneficial in a future project to form difference rasters for all 19 bioclimatic variables. Figures 2-3 also do suggest that another variables, such as elevation, may be important to include due to the indication of higher presence in the Champlain Valley (close to Lake Champlain) and the banks of the Connecticut River. Elevation is used in many SDMs (Hof et al. 2012), so its introduction to a future project would not be out of the question. For now, it is able to be surmised that the climatic conditions of southern New England, based on 19 variables which are designed to be relevant to species distributions, may become insufficient for tree-of-heaven in the coming 40 years; however, the opposite becomes true for much of northern New England (Figures 2-3). Extracting variable importance, as is possible for random forest, would be helpful in determining exactly which bioclimatic variables exert the most influence over the model and thus have the most predictive power.

Potential importance

Invasive plant species watch lists, a topic discussed by Reaser et al. (2020b) and Jarnevich et al. (2023), may benefit from the results in that land which is potentially historically suitable for tree-of-heaven is identified by Figure 1. For example, an organization with land in eastern Connecticut which has not seen tree-of-heaven on its property may be interested in the suggestion that its land is still suitable for the species to be found according to a statistical model.

Additionally, land managers may be interested in incorporating predictions into short- or mid-term planning for invasive species management as part of existing efforts to set out plans for how their organization may approach land management in future decades. Land managers near Burlington, VT, along the Connecticut River, and across central Maine may benefit from noting down that the species could potentially be found in their land, given sufficient dispersal, in the coming decades. As of writing this document, the first future period has begun, and its end is just 15 years away. Due to this, it could be important that land managers in these areas learn more about the ecology of tree-of-heaven and how scientists have opted to remove it in the past, which could include biological control measures (Harris et al. 2013).

Future work

This project is well-suited to expansion. In addition to the extensions discussed in the last part of the “current vs. future range” section, it could be continued in various ways: for one, it is perhaps conspicuous that the highest-resolution version of the Worldclim data (30 seconds) was not used to account for space constraints, computational time, and the apparent failure of the `cimp6_tile` function, which would allow for the download of data which has already been cropped compared to a global extent. One may have higher confidence in a model which takes advantage of said resolution, especially due to the `ENMevaluate` warning message referring to many of the points being in the same grid cell; this issue may be solved by having a smaller resolution, for which there is only 0.5 minutes (30 seconds). Even if the 30 arcsecond resolution is infeasible to use in a future project, comparison of models built upon the two other resolutions (10 and 5 minutes) is still a potentially valuable avenue for model validation (Özdemir 2024).

Additionally, different models which are used for SDMs, such as random forest (e.g., Evans et al. 2010) could be compared with the MaxEnt results to account for the idea that MaxEnt is only one possible way to model the presence of species. Comparison between MaxEnt and random forest is suggested as appropriate under certain circumstances in a tutorial article by Esri, the creator of the popular ArcGIS family of software (Nieto and Liu 2022). ArcGIS Pro itself does have an implementation of MaxEnt (Liu 2022), and it is also perhaps worth comparing rasters created in R and ArcGIS using identical parameters in order to investigate whether computational environment affects the prediction surfaces.

Last, it would be beneficial to examine whether a similar range expansion is predicted by MaxEnt to occur for other species which are invasive in Massachusetts. Such a project would serve to augment hypothetical watch lists (after Jarnevich et al. 2023) for northern land managers who may not have had to consider the threat of more southern species. An approach testing all 36 invasive species in Massachusetts under similar test conditions, while time-consuming (high values of k-fold validation for `ENMevaluate` can take upwards of 30 minutes) and computationally intensive, would be an intriguing comparison. Methods already implemented to some degree in this script, such as using `rgbif` to remotely download GBIF data, may save a small amount of time on manual menu navigation, however. Tree-of-heaven serves as a surrogate to explain to a land manager the threat of climate change exacerbating the problem of invasive species, but this message can be enhanced by increasing the sample size of species who have the potential to move northward.

Acknowledgements

I extend my gratitude to Instructor James Tsalah for the reference materials, particularly related to the use of the functions `ENMevaluate` and `dismo::maxent`.

Bibliography

Ahmed, N., Atzberger, C., & Zewdie, W. (2020). Integration of remote sensing and bioclimatic data for prediction of invasive species distribution in data-poor regions: A review on challenges and opportunities. Environmental Systems Research, 9(1), 32. https://doi.org/10.1186/s40068-020-00195-0 (https://doi.org/10.1186/s40068-020-00195-0).

Brooks, R. K., Barney, J. N., & Salom, S. M. (2021). The invasive tree, Ailanthus altissima, impacts understory nativity, not seedbank nativity. Forest Ecology and Management, 489, 119025. https://doi.org/10.1016/j.foreco.2021.119025 (https://doi.org/10.1016/j.foreco.2021.119025).

Evans, J. S., Murphy, M. A., Holden, Z. A., & Cushman, S. A. (2011). Modeling Species Distribution and Change Using Random Forest. In C. A. Drew, Y. F. Wiersma, & F. Huettmann (Eds.), Predictive Species and Habitat Modeling in Landscape Ecology: Concepts and Applications (pp. 139–159). Springer. https://doi.org/10.1007/978-1-4419-7390-0_8 (https://doi.org/10.1007/978-1-4419-7390-0_8).

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology, 37(12), 4302–4315. https://doi.org/10.1002/joc.5086 (https://doi.org/10.1002/joc.5086).

GBIF: Global Biodiversity Information Facility. (2025, May 18). Ailanthus altissima (Mill.) Swingle. https://www.gbif.org/species/3190653 (https://www.gbif.org/species/3190653)

Harris, P. T., Cannon, G. H., Smith, N. E., & Muth, N. Z. (2013). Assessment of plant community restoration following Tree-of-Heaven (Ailanthus altissima) control by Verticillium albo-atrum. Biological Invasions, 15(9), 1887–1893. https://doi.org/10.1007/s10530-013-0430-2 (https://doi.org/10.1007/s10530-013-0430-2).

Hess, M. C. M., Mesléard, F., & Buisson, E. (2019). Priority effects: Emerging principles for invasive plant species management. Ecological Engineering, 127, 48–57. https://doi.org/10.1016/j.ecoleng.2018.11.011 (https://doi.org/10.1016/j.ecoleng.2018.11.011).

Hof, A. R., Jansson, R., & Nilsson, C. (2012). The usefulness of elevation as a predictor variable in species distribution modelling. Ecological Modelling, 246, 86–90. https://doi.org/10.1016/j.ecolmodel.2012.07.028 (https://doi.org/10.1016/j.ecolmodel.2012.07.028).

Janowiak, M. K., D’Amato, A. W., Swanston, C. W., Iverson, L., Thompson, F. R., Dijak, W. D., Matthews, S., Peters, M. P., Prasad, A., Fraser, J. S., Brandt, L. A., Butler-Leopold, P., Handler, S. D., Shannon, P. D., Burbank, D., Campbell, J., Cogbill, C., Duveneck, M. J., Emery, M. R., … Templer, P. H. (2018). New England and northern New York forest ecosystem vulnerability assessment and synthesis: A report from the New England Climate Change Response Framework project. Gen. Tech. Rep. NRS-173. Newtown Square, PA: U.S. Department of Agriculture, Forest Service, Northern Research Station. 234 p., 173, 1–234. https://doi.org/10.2737/nrs-gtr-173 (https://doi.org/10.2737/nrs-gtr-173).

Jarnevich, C., Engelstad, P., LaRoe, J., Hays, B., Hogan, T., Jirak, J., Pearse, I., Prevéy, J., Sieracki, J., Simpson, A., Wenick, J., Young, N., & Sofaer, H. R. (2023). Invaders at the doorstep: Using species distribution modeling to enhance invasive plant watch lists. Ecological Informatics, 75, 101997. https://doi.org/10.1016/j.ecoinf.2023.101997 (https://doi.org/10.1016/j.ecoinf.2023.101997).

Lake, T. A., Briscoe Runquist, R. D., & Moeller, D. A. (2020). Predicting range expansion of invasive species: Pitfalls and best practices for obtaining biologically realistic projections. Diversity and Distributions, 26(12), 1767–1779. https://doi.org/10.1111/ddi.13161 (https://doi.org/10.1111/ddi.13161) Larson, D. L., Phillips-Mao, L., Quiram, G., Sharpe, L., Stark, R., Sugita, S., & Weiler, A. (2011). A framework for sustainable invasive species management: Environmental, social and economic objectives. Journal of Environmental Management, 92(1), 14–22. https://doi.org/10.1016/j.jenvman.2010.08.025 (https://doi.org/10.1016/j.jenvman.2010.08.025).

Mainali, K. P., Warren, D. L., Dhileepan, K., McConnachie, A., Strathie, L., Hassan, G., Karki, D., Shrestha, B. B., & Parmesan, C. (2015). Projecting future expansion of invasive species: Comparing and improving methodologies for species distribution modeling. Global Change Biology, 21(12), 4464–4480. https://doi.org/10.1111/gcb.13038 (https://doi.org/10.1111/gcb.13038).

McAvoy, T. J., Snyder, A. L., Johnson, N., Salom, S. M., & Kok, L. T. (2012). Road Survey of the Invasive Tree-of-Heaven (Ailanthus altissima) in Virginia. Invasive Plant Science and Management, 5(4), 506–512. https://doi.org/10.1614/IPSM-D-12-00039.1 (https://doi.org/10.1614/IPSM-D-12-00039.1).

Meinshausen, M., Nicholls, Z. R. J., Lewis, J., Gidden, M. J., Vogel, E., Freund, M., Beyerle, U., Gessner, C., Nauels, A., Bauer, N., Canadell, J. G., Daniel, J. S., John, A., Krummel, P. B., Luderer, G., Meinshausen, N., Montzka, S. A., Rayner, P. J., Reimann, S., … Wang, R. H. J. (2020). The shared socio-economic pathway (SSP) greenhouse gas concentrations and their extensions to 2500. Geoscientific Model Development, 13(8), 3571–3605. https://doi.org/10.5194/gmd-13-3571-2020 (https://doi.org/10.5194/gmd-13-3571-2020).

Miller, J. (2010). Species Distribution Modeling. Geography Compass, 4(6), 490–509. https://doi.org/10.1111/j.1749-8198.2010.00351.x (https://doi.org/10.1111/j.1749-8198.2010.00351.x).

MIPAG: Massachusetts Invasive Plant Advisory Group. (2022). Plants voted as: INVASIVE. https://www.massnrc.org/mipag/invasive.htm (https://www.massnrc.org/mipag/invasive.htm).

Nieto, A., & Liu, J. (2022, May 4). Eight tips to help you make better presence prediction models with Presence-only Prediction (MaxEnt). ArcGIS Blog. https://www.esri.com/arcgis-blog/products/arcgis-pro/analytics/eight-tips-to-help-you-make-better-presence-prediction-models-with-presence-only-prediction-maxent (https://www.esri.com/arcgis-blog/products/arcgis-pro/analytics/eight-tips-to-help-you-make-better-presence-prediction-models-with-presence-only-prediction-maxent).

NISIC: National Invasive Species Information Center. (n.d.). Early Detection and Rapid Response. https://www.invasivespeciesinfo.gov/subject/early-detection-and-rapid-response (https://www.invasivespeciesinfo.gov/subject/early-detection-and-rapid-response).

Özdemir, S. (2024). Testing the Effect of Resolution on Species Distribution Models Using Two Invasive Species. Polish Journal of Environmental Studies, 33(2), 1325–1335. https://doi.org/10.15244/pjoes/166353 (https://doi.org/10.15244/pjoes/166353).

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. Ecological Modelling, 190(3), 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026 (https://doi.org/10.1016/j.ecolmodel.2005.03.026).

Reaser, J. K., Burgiel, S. W., Kirkey, J., Brantley, K. A., Veatch, S. D., & Burgos-Rodríguez, J. (2020). The early detection of and rapid response (EDRR) to invasive species: A conceptual framework and federal capacities assessment. Biological Invasions, 22(1), 1–19. https://doi.org/10.1007/s10530-019-02156-w (https://doi.org/10.1007/s10530-019-02156-w).

Reaser, J. K., Frey, M., & Meyers, N. M. (2020). Invasive species watch lists: Guidance for development, communication, and application. Biological Invasions, 22(1), 47–51. https://doi.org/10.1007/s10530-019-02176-6 (https://doi.org/10.1007/s10530-019-02176-6).

Soler, J., & Izquierdo, J. (2024). The Invasive Ailanthus altissima: A Biology, Ecology, and Control Review. Plants, 13(7), Article 7. https://doi.org/10.3390/plants13070931 (https://doi.org/10.3390/plants13070931).

Srivastava, V., Lafond, V., & Griess, V. C. (2019). Species distribution models (SDM): Applications, benefits and challenges in invasive species management. CABI Reviews, 2019, 1–13. https://doi.org/10.1079/PAVSNNR201914020 (https://doi.org/10.1079/PAVSNNR201914020).

Tang, G., Beckage, B., & Smith, B. (2012). The potential transient dynamics of forests in New England under historical and projected future climate change. Climatic Change, 114(2), 357–377. https://doi.org/10.1007/s10584-012-0404-x (https://doi.org/10.1007/s10584-012-0404-x).

Wimberly, M. C. (2023). Chapter 12 – Application—Species Distribution Modeling. In Geographic Data Science with R: Visualizing and Analyzing Environmental Change. https://bookdown.org/mcwimberly/gdswr-book/application---species-distribution-modeling.html#climate-change-projections (https://bookdown.org/mcwimberly/gdswr-book/application---species-distribution-modeling.html#climate-change-projections)