Dotted duckweed maxent study: Preliminary model run (rmd file #1)

Debora

2024-05-02

Purpose: To create a preliminary model using pre-selected bioclimatic variables. I will check for correlations and importance of each variable to the model to perform a subsequent variable selection which will include the most important variables with the lowest levels of correlation between them.

Initial selection of bioclimatic variables

- Reflect thermal dependency of duckweeds in terms of population growth
- Account for the dependency on moderate precipitation

Load lake water temperature and air temperature rasters

```
# Lake water temperature raster based on satellite measurements (Armitage, 2023; https://onlinelibrary.
options(timeout=600)
url <- "https://datadryad.org/stash/downloads/file_stream/1895801"
temp_file <- tempfile()
temp_unzipped1 <- tempfile()
temp_unzipped2 <- tempfile()
download.file(url, destfile = temp_file, mode="wb")
unzip(temp_file, exdir = temp_unzipped1)
unzip(paste0(temp_unzipped1,"/LakeTemps_Code/rasters.zip"), exdir = temp_unzipped2)
lake <- raster::brick(paste0(temp_unzipped2,"/rasters/bioclim_lakes_10km.tif"))

air <- geodata::worldclim_global(var = 'bio', res = 5, download = T, path = 'data')
air <- as(air, "Raster")
air <- brick(air)
```

Obtain occurrence records from GBIF and remove duplicates and records missing information

Downloaded from GBIF on Mar 28, 2024 https://doi.org/10.15468/dl.7uqs9k

```
temp <- tempfile()
download.file("https://api.gbif.org/v1/occurrence/download/request/0049626-240321170329656.zip",temp)
lp <- read.csv(unz(temp, "occurrence.txt"), head = TRUE, sep="\t")
```

Clean up dataset

(removing data without coordinates and duplicates)

```
# relabel latitude and longitude columns
colnames(lp)[c(98,99)] = c("lat","lon")
```

```
# removing data without geographic coordinates
lp <- subset(lp, !is.na(lon) & !is.na(lat))

# removing duplicates
lp_clean  <-  lp[!duplicated(lp[c("lat","lon")]),]
```

Add occurrences from literature

- The references for each of these coordinates are in Appendix 3
- For cases where coordinates are not exact, coordinates of a given country's capital were used

```
# exact records obtained from published papers
europe_lit <- data.frame(lon=c(4.51,4.95,4.31,4.25,6.12,5.23,12.27,12.45,25.15,12.97),
                         lat=c(52.18,52.23,52,52,51.31,51.48,43.56,41.9,60.27,56.25))

# inexact records obtained from published papers
europe_lit_unreported <- data.frame(lon=c(4.47,0.11,12.57,12.57,6.14),
                                    lat=c(50.5,51.5,41.87,41.87,46.2))

lp_clean <- lp_clean[c("lon","lat")]
lp_clean_all <- rbind(lp_clean, europe_lit, europe_lit_unreported)
```

Remove observations landing on NA predictor values

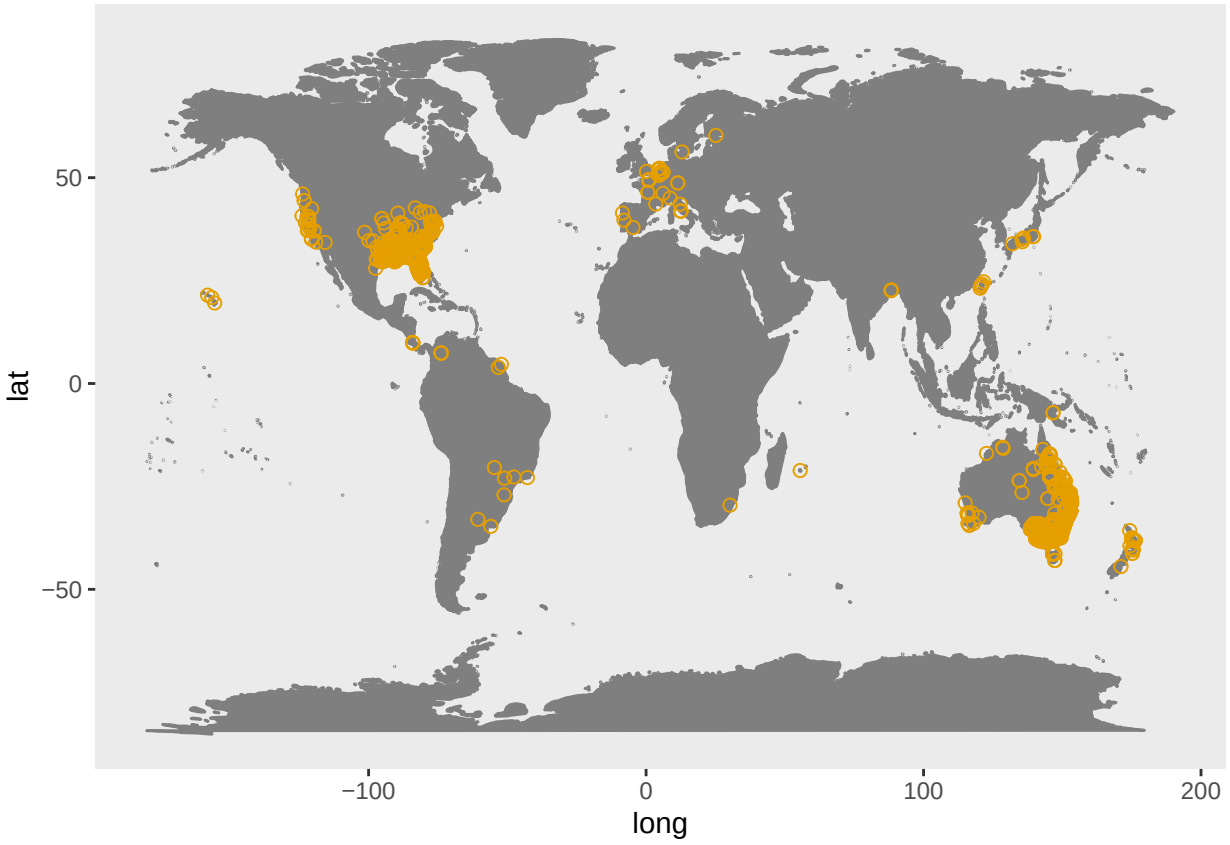
(based on lake raster, which has less observations)

```
v <- raster::extract(lake_crop, lp_clean_all)

# which points have NA values?
i <- which(apply(is.na(v), 1,  sum) > 0)

# remove these from dataset
lp_clean_all <- lp_clean_all[-i, ]
```

Create datasets for each geographic location

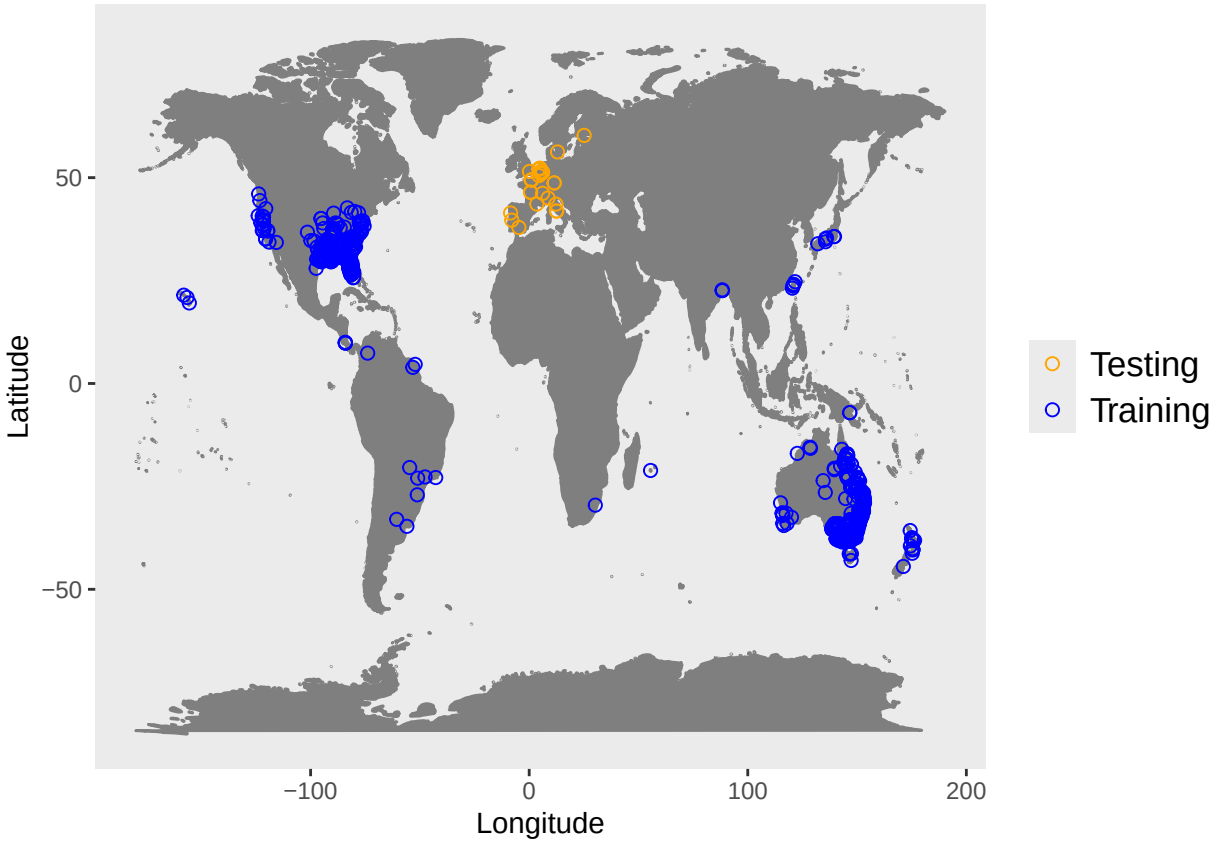
```
#subsetting by geographical region
world <- lp_clean_all [!(lp_clean_all$lon >= -50 & lp_clean_all$lon <= 66
                         & lp_clean_all$lat >= 20 & lp_clean_all$lat <= 72.01), ]

europe <- lp_clean_all [(lp_clean_all$lon >= -50 & lp_clean_all$lon <= 66
                         & lp_clean_all$lat >= 20 & lp_clean_all$lat <= 72.01), ]
```

Plot cleaned dataset of occurrences (GBIF and literature records)



Perform spatial thinning to reduce spatial autocorrelation

```
# transform dataframe into spatial points dataframe
coordinates(world) <- ~lon+lat
coordinates(europe) <- ~lon+lat

set.seed(2018)
# to reduce sampling bias (disproportional reporting in some areas), a single observation was sampled u

world_sampled <- gridSample(world, lake_crop, n=1) # sample 1 observation per area
world_sampled <- world_sampled[!duplicated(world_sampled), ]

europe_sampled <- gridSample(europe, lake_crop, n=2) # sample 2 observations per area
# this is the minimum sampling allowing for a final testing dataset > 30 observations
europe_sampled <- europe_sampled[!duplicated(europe_sampled), ]
```



Relabel dataset as training (world excluding Europe) and testing (Europe)

```
pres_train <- world_sampled
pres_test <- europe_sampled
```

Create study area, sample background points, and run models

```
raster_list <- list(lake_crop,air_crop)
backg_train_list <- list()
maxent_output_list <- list()
best_aic_list <- list()
best_auc_list <- list()
best_model_list <- list()
importance_list <- list()
correl_list <- list()

count = 1
for (i in raster_list){
# Crop environment to correspond to species distribution range (+/- 2 degrees)
model.extent<-extent(min(pres_train$lon)-10,max(pres_train$lon)+10,
                     min(pres_train$lat)-10,max(pres_train$lat)+10)

crop_raster <- crop(i,model.extent)

rm(lake,air,lake_crop,air_crop,stk)
```

```
occ_buff <- buffer(pres_train, 600000, dissolve=TRUE) # width parameter = 600000m (600km)
# Reference for selecting buffer extent:
# https://www.sciencedirect.com/science/article/pii/S0304380023001850

# crop study area to buffer extent
studyArea <- crop(crop_raster, extent(occ_buff))

# mask the non buffer areas
studyArea <- mask(studyArea, occ_buff)
# output will still be a raster stack, just of the study area

# Randomly sample points
# Sample same number as our observed points inside the buffer
# to create background points, or hypothetical areas where species
# could either be found or not

set.seed(2022)
backg_train <- randomPoints(studyArea, n=length(pres_train$lon), p=pres_train, extf=1)
colnames(backg_train) <- c("lon","lat")

backg_train_list[[count]] <- backg_train

# Run candidate models
e.mx.l <- ENMevaluate(occs = pres_train, envs = crop_raster, bg = backg_train,
                                algorithm = 'maxent.jar', partitions = 'block',
                                tune.args = list(fc = c("L","LQ"),
                                rm = seq(0.5, 4, by = 0.5)))

maxent_output_list[[count]] <- e.mx.l

  result <- eval.results(e.mx.l)

  best_aic <- result[which.min(result$AICc),]
  best_auc <- result[which.max(result$Mean.testing.AUC),]

  model <- eval.models(e.mx.l)[[best_aic$tune.args]]
  best_model_list[[count]] <- model

  best_aic_list[[count]] = as.data.frame(best_aic)
  best_auc_list[[count]] = as.data.frame(best_auc)

  importance <- eval.variable.importance(e.mx.l)[[best_aic$tune.args]]

  importance_list[[count]] <- importance

  correl <- ENMTools::raster.cor.matrix(crop_raster, method = "pearson")

  correl_list[[count]] <- correl

count=count+1
}
```

```
## Warning in .couldBeLonLat(x): CRS is NA. Assuming it is longitude/latitude
```

```
## Package ecospat is not installed, so Continuous Boyce Index (CBI) cannot be calculated.
## *** Running initial checks... ***
## * Removed 1 occurrence localities that shared the same grid cell.
## * Clamping predictor variable rasters...
## * Model evaluations with spatial block (4-fold) cross validation and lat_lon orientation...
##
## *** Running ENMeval v2.0.4 with maxent.jar v3.4.3 from dismo package v1.3.14 ***
##   |                                                                              |
## ENMevaluate completed in 11 minutes 18.2 seconds.
## Warning in rm(lake, air, lake_crop, air_crop, stk): object 'lake' not found
## Warning in rm(lake, air, lake_crop, air_crop, stk): object 'air' not found
## Warning in rm(lake, air, lake_crop, air_crop, stk): object 'lake_crop' not
## found
## Warning in rm(lake, air, lake_crop, air_crop, stk): object 'air_crop' not found
## Warning in rm(lake, air, lake_crop, air_crop, stk): object 'stk' not found
## Warning in .couldBeLonLat(x): CRS is NA. Assuming it is longitude/latitude
## Package ecospat is not installed, so Continuous Boyce Index (CBI) cannot be calculated.
## *** Running initial checks... ***
## * Removed 1 occurrence localities that shared the same grid cell.
## * Clamping predictor variable rasters...
## * Model evaluations with spatial block (4-fold) cross validation and lat_lon orientation...
##
## *** Running ENMeval v2.0.4 with maxent.jar v3.4.3 from dismo package v1.3.14 ***
##   |                                                                              |
## ENMevaluate completed in 12 minutes 4.1 seconds.
```

Select variables which contribute the most and have the lowest correlations

Criteria: - BIO1 is always selected (for response curve comparison with thermal performance) - At least one variable related to precipitation is selected - cutoff: permutation importance >5%

Table 1: Lake temperature preliminary model

	variable	percent.contribution	permutation.importance
5	bioclim_lakes_10km_15	40.89	7.44
10	bioclim_lakes_10km_7	22.92	30.17
3	bioclim_lakes_10km_11	10.93	23.31
9	bioclim_lakes_10km_3	10.05	2.55
2	bioclim_lakes_10km_10	4.53	13.81
8	bioclim_lakes_10km_2	3.72	2.77
1	bioclim_lakes_10km_1	2.84	16.53
4	bioclim_lakes_10km_12	2.24	1.97
7	bioclim_lakes_10km_18	1.18	0.05
6	bioclim_lakes_10km_17	0.69	1.41

	variable	percent.contribution	permutation.importance

Table 2: Air temperature preliminary model

	variable	percent.contribution	permutation.importance
5	wc2.1_5m_bio_15	32.95	2.25
10	wc2.1_5m_bio_7	28.43	29.04
3	wc2.1_5m_bio_11	21.26	42.02
1	wc2.1_5m_bio_1	6.26	0.03
9	wc2.1_5m_bio_3	3.86	5.82
2	wc2.1_5m_bio_10	3.77	11.04
8	wc2.1_5m_bio_2	2.46	8.31
6	wc2.1_5m_bio_17	0.75	1.37
4	wc2.1_5m_bio_12	0.18	0.00
7	wc2.1_5m_bio_18	0.08	0.12

- cutoff: correlation coefficient <0.8 (Elith et al. 2010)

Table 3: Correlation matrix: lake-based bioclimatic variables

	BIO15	BIO7	BIO11	BIO3	BIO10	BIO2	BIO1	BIO12	BIO18	BIO17
BIO15	1.00	-0.13	0.22	0.06	0.21	0.02	0.24	-0.18	0.02	-0.50
BIO7		1.00	-0.84	-0.82	-0.32	0.22	-0.69	-0.55	-0.42	-0.29
BIO11			1.00	0.81	0.76	0.04	0.97	0.43	0.30	0.16
BIO3				1.00	0.40	0.24	0.69	0.56	0.41	0.30
BIO10					1.00	0.15	0.90	0.10	0.03	-0.04
BIO2						1.00	0.08	-0.12	-0.09	-0.14
BIO1							1.00	0.32	0.21	0.08
BIO12								1.00	0.84	0.76
BIO18									1.00	0.51
BIO17										1.00

Table 4: Correlation matrix: air-based bioclimatic variables

	BIO15	BIO7	BIO11	BIO3	BIO10	BIO2	BIO1	BIO12	BIO18	BIO17
BIO15	1.00	0.00	0.20	0.13	0.27	0.40	0.25	-0.22	-0.13	-0.55
BIO7		1.00	-0.86	-0.87	-0.32	0.37	-0.72	-0.62	-0.39	-0.37
BIO11			1.00	0.83	0.74	-0.01	0.97	0.40	0.16	0.13
BIO3				1.00	0.37	-0.04	0.72	0.56	0.31	0.31
BIO10					1.00	0.23	0.88	0.01	-0.13	-0.11
BIO2						1.00	0.08	-0.56	-0.48	-0.50
BIO1							1.00	0.29	0.08	0.05
BIO12								1.00	0.78	0.74
BIO18									1.00	0.56
BIO17										1.00