

Informing Malleefowl Recovery Through Species Distribution Modelling

Michael Jordan

2023-12-01



Figure 1: Malleefowl; image from Macaulay Library at the Cornell Lab (ML82227771).

Introduction & Rationale

The malleefowl (*Leipoa ocellata*) is a large (~2kg) ground-dwelling bird endemic to Australia. It is a member of the Megapodiidae family, which incubate their eggs in mounds that provide an external source of heat rather than in nests using body heat (Parsons, Short & Roberts, 2008). Within the Megapodiidae, the malleefowl is notable for being the only species to inhabit semi-arid environments rather than damp forests and for having developed a particularly elaborate mound building technique (Benshemesh, 2007).

Malleefowl mounds are large (~3-5m in diameter x 1m deep), scraped out of the ground, filled with leaf litter, and covered with sand, which acts as insulation (Benshemesh, 2007). Within mounds, solar radiation and

decomposition of organic matter maintain a regular temperature within 3-5°C of 34°C, which is necessary for egg development (Booth, 1986). Because of this intriguing incubation strategy, malleefowl have become locally and internationally beloved and a significant draw for tourists (Benshemesh, 2007). Unfortunately, although malleefowl were widely distributed across Australia prior to European colonization (Parsons, Short & Roberts, 2008), today malleefowl are listed as vulnerable on the IUCN Red List and under the Australian Environment Protection and Biodiversity Conservation Act of 1999.

In 2007, to correct malleefowl population declines, the Australian government instituted a \$2.4M AUD malleefowl National Recovery Plan (NRP) (Benshemesh, 2007). To justify this investment, the NRP cited steep declines in monitored malleefowl populations. It set several objectives, including reducing habitat loss and reducing grazing pressure by livestock. However, the NRP may not succeed if it fails to understand current and future malleefowl distribution. Although some current malleefowl populations are monitored, malleefowl distribution is not well established across all of Australia (Benshemesh, 2007). Additionally, climate change may drive changes to malleefowl habitat. Within monitored areas, malleefowl population concentration is observed to be highest in areas with high rainfall, likely due to increased food supply (Benshemesh, 2007). Prior research also indicates that temperature may be significant, since egg temperature effects breeding success (Stenhouse & Moseby, 2022). If rainfall or temperature shift, malleefowl may gain or lose viable habitat.

This report seeks to inform the malleefowl NRP by modeling current and future malleefowl distribution using environmental and climatic variables. This information will allow for better targeting of conservation funding in areas of greatest long term potential. Additionally, this reports conducts statistical analysis on the Megapodiidae family to determine if dispersal ability relates to extinction risk.

Modelling Current and Future Malleefowl Species Distribution

Analysis was completed using R version 4.3.1. Malleefowl occurrence data were downloaded from the Global Biodiversity Information Facility (GBIF, 2022) database using the `dismo` package.

```
# Load necessary libraries
library(dismo)
library(terra)
library(geodata)
library(sf)
library(rnaturalearth)
library(dplyr)
library(bestglm)
library(ggplot2)
library(kableExtra)
library(scales)

# Import data from GBIF and retain only relevant columns
m.full <- dismo::gbif("Leipoa", "ocellata")
keep <- c("adm1", "adm2", "basisOfRecord", "cloc", "collectionCode",
          "coordinateUncertaintyInMeters", "day", "eventDate", "eventTime",
          "gbifID", "geodeticDatum", "georeferenceVerificationStatus",
          "higherGeography", "identifiedBy", "informationWithheld",
          "institutionCode", "lat", "locality", "lon", "month", "occurrenceID",
          "occurrenceRemarks", "protocol", "recordedBy", "year")
m.e <- m.full[, names(m.full) %in% keep]
```

Data were cleaned to include only trustworthy, non-duplicate observations from years used as a historic baseline for climate forecasting. Geo coordinate uncertainty was tolerated up to 1km, since this provided

reasonable precision relative to GIS data resolution (~3.5-4.5km). Data were checked for incorrect geo coordinate zeroing, but none was found.

```
# Clean data: 1) filter for observation occurrences, 2) remove occurrences
# without geo coords, 3) remove uncertain occurrences (unverified, obfuscated
# geo coords, > 1km geo coord uncertainty), 3) filter for records in baseline
# historical years
m.e <- m.e[
  m.e$basisOfRecord %in% c("HUMAN_OBSERVATION", "MACHINE_OBSERVATION")
  & !is.na(m.e$lon)
  & !is.na(m.e$lat)
  & (
    is.na(m.e$georeferenceVerificationStatus)
    | m.e$georeferenceVerificationStatus == "verified")
  & is.na(m.e$informationWithheld)
  & (
    is.na(m.e$coordinateUncertaintyInMeters)
    | m.e$coordinateUncertaintyInMeters <= 1000)
  & !is.na(m.e$year)
  & m.e$year >= 1970
  & m.e$year <= 2000, ]

# Drop observations that are duplicative at the day/lat/lon level.
m.e$dup <- duplicated(m.e[, c("eventDate", "lon", "lat")])
m.e <- m.e[m.e$dup == FALSE, ]

# Create an object with only geo coord columns
m <- m.e[, c(17, 19)]
```

Overall, data cleaning reduced records from 10,501 to 731. Observations were plotted using the terra and sf packages to validate (Figure 2).

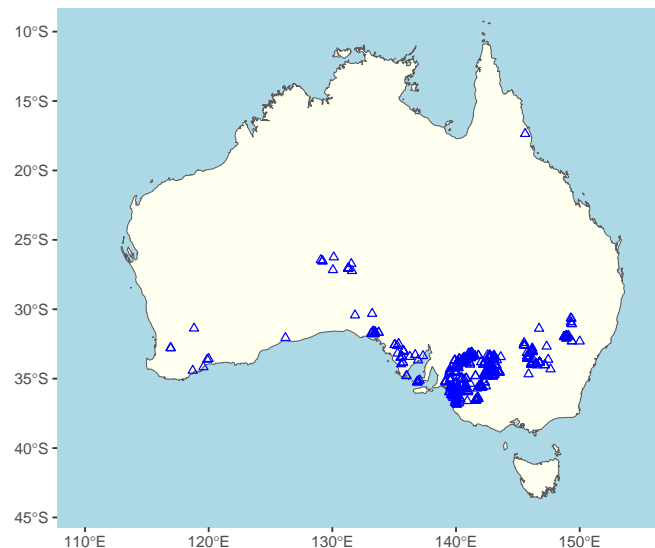


Figure 2: Malleefowl observations 1970-2000.

To estimate malleefowl distribution, historical (1970-2000) and future climate data were pulled from WorldClim 2.1 using the geodata package. Future climate data were pulled for two sequential time periods

(2021-2040, 2041-2060) using the Hadley GEM3 model and Shared Socioeconomic Pathway 2-4.5.

These future time periods were selected in order to evaluate the time scale of any malleefowl distribution changes. SSP 2-4.5 was selected because it represents a “middle ground” scenario and so should provide a conservative estimate of malleefowl distribution changes.

The WorldClim data provided 19 climate measurements, many of which were similar or lacked a plausible biological mechanism to impact malleefowl distribution. The following climatic variables were selected as potential model coefficients based on their being distinct and biologically relevant (Table 1).

Table 1: Climate variable definitions.

Climate Variable	Definition
Bio1	Annual Mean Temperature
Bio3	Isothermality
Bio5	Max Temperature of Warmest Month
Bio6	Min Temperature of Coldest Month
Bio12	Annual Precipitation
Bio13	Precipitation of Wettest Month
Bio14	Precipitation of Driest Month

Because malleefowl have preferences based on elevation, sandiness of soil, and fertility of soil (Benshemesh, 2007), the geodata package was used to pull elevation data from the Shuttle Radar Topography Mission and soil composition from the SoilGRIDS database. Sand composition was measured as the percentage of $> .05\text{mm}$ particles. Soil fertility was measured using organic carbon density (OCD) in kg/m^3 .

Climate and elevation data were pulled at a 2.5 arc-minute resolution. Soil data were pulled at a 30 arc-second resolution and aggregated to 2.5 arc-minutes. A 2.5 arc-minute resolution was selected because it is reasonably granular, but encompasses an individual malleefowl’s territory. At Australian latitudes, 2.5 arc-minutes is $\sim 2.2\text{-}2.8$ miles, whereas radio-tracking studies have demonstrated that malleefowl annual range is within a few miles (Benshemesh, 2007).

```
# Pull climate, elevation and soil data and crop to Australia geo extent
bioclim.h <- geodata::worldclim_global(var = "bio", res = 2.5, path = "data")
bioclim.h <- terra::crop(bioclim.h, ext)
bioclim.f1 <- geodata::cmip6_world(
  model = "HadGEM3-GC31-LL",
  ssp = "585",
  time = "2021-2040",
  var = "bioc",
  path = "data",
  res = 2.5)
bioclim.f1 <- terra::crop(bioclim.f1, ext)
bioclim.f2 <- geodata::cmip6_world(
  model = "HadGEM3-GC31-LL",
  ssp = "585",
  time = "2041-2060",
  var = "bioc",
  path = "data",
  res = 2.5)
bioclim.f2 <- terra::crop(bioclim.f2, ext)
elev <- geodata::elevation_global(res = 2.5, path = "data")
elev <- terra::crop(elev, ext)
sand <- geodata::soil_world(
  var = "sand",
```

```

stat = "mean",
depth = 60,
path = "data")
sand <- terra::crop(sand, ext)
carbon <- geodata::soil_world(
  var = "ocd",
  stat = "mean",
  depth = 60,
  path = "data")
carbon <- terra::crop(carbon, ext)

# Aggregate soil rasters from 30 arc-second to 2.5 arc-minute resolution
sand <- terra::aggregate(sand, fact = 5, fun = mean)
carbon <- terra::aggregate(carbon, fact = 5, fun = mean)

# Align and simplify column names
bioclim.names <- paste0("bio", 1:19)
names(bioclim.h) <- bioclim.names
names(bioclim.f1) <- bioclim.names
names(bioclim.f2) <- bioclim.names
names(elev) <- c("elev")
names(sand) <- c("sand")
names(carbon) <- c("ocd")

# Select subset of variables relevant to test in a model
bioclim.h <- terra::subset(
  bioclim.h,
  c("bio1", "bio3", "bio5", "bio6", "bio12", "bio13", "bio14"))
bioclim.f1 <- terra::subset(
  bioclim.f1,
  c("bio1", "bio3", "bio5", "bio6", "bio12", "bio13", "bio14"))
bioclim.f2 <- terra::subset(
  bioclim.f2,
  c("bio1", "bio3", "bio5", "bio6", "bio12", "bio13", "bio14"))

```

For projection, Australian Albers was selected because it is an equal area projection, enabling accurate area comparisons across latitudes.

Climate, elevation, and soil rasters were projected onto Australian Albers and combined.

```

# Create a grid and project objects onto the Australian Albers projection
ntm.grid <- terra::rast(
  ext(-2740000, 2400000, -5252000, -944000),
  res = 20000,
  crs = "EPSG:3577")
m.sf <- sf::st_transform(m.sf, crs = "EPSG:3577")
bioclim.h <- terra::project(bioclim.h, ntm.grid)
bioclim.f1 <- terra::project(bioclim.f1, ntm.grid)
bioclim.f2 <- terra::project(bioclim.f2, ntm.grid)
elev <- terra::project(elev, ntm.grid)
sand <- terra::project(sand, ntm.grid)
carbon <- terra::project(carbon, ntm.grid)

# Add elevation, sand, and carbon data to historical and future rasters

```

```

bioclim.h <- c(bioclim.h, elev, sand, carbon)
bioclim.f1 <- c(bioclim.f1, elev, sand, carbon)
bioclim.f2 <- c(bioclim.f2, elev, sand, carbon)

# Use au vector from rnaturalearth to mask non-AU islands
au.msk <- sf::st_transform(au, crs = "EPSG:3577")
bioclim.h <- terra::mask(bioclim.h, au.msk["type"])
bioclim.f1 <- terra::mask(bioclim.f1, au.msk["type"])
bioclim.f2 <- terra::mask(bioclim.f2, au.msk["type"])

```

Logistic Generalized Linear Model (GLM) regression was chosen as a modelling technique, since it enables modelling the probability of an event. Since logistic regression requires both presence and absence data, background psuedo-absence points equal to the observation count ($n = 731$) were generated at random (excluding observation points) using the dismo package.

```

# Create random psuedo absence geo coords
n.pseudo <- nrow(m.sf)
land <- bioclim.h[["elev"]] > 0
pseudo.dismo <- dismo::randomPoints(
  mask = as(land, "Raster"),
  n = n.pseudo,
  p = sf::st_coordinates(m.sf))
pseudo.dismo <- sf::st_as_sf(
  data.frame(pseudo.dismo),
  coords = c("x", "y"),
  crs = 3577)

```

Observation and pseudo-absence points were visually plotted to inspect for errors (Figure 3).

```

# Plot presence and psuedo-absence points to visually inspect for errors
par(mar = c(1, 1, 1, 1))
terra::plot(land, col = "grey", legend = FALSE, axes = FALSE)
terra::plot(sf::st_geometry(m.sf), add = TRUE, col = "blue", cex = 1, pch = 2)
terra::plot(pseudo.dismo, add = TRUE, col = "black", cex = .2, pch = 19)
legend(
  "bottomleft",
  legend = c("Observation", "Pseudo-Absence"),
  col = c("blue", "black"),
  pch = c(2, 19),
  pt.cex = c(1, .2),
  cex = 1,
  inset = c(.2, .03))

```

One potential problem with this approach is that malleefowl are rare and not exhaustively surveyed. Therefore, some psuedo-absence points may diverge from true absence (i.e., contain malleefowl). If this is the case, model estimates will be biased.

To reserve 20% of data for model evaluation, the dismo package was used to add quintiles. Data were combined into a single presence/absence special feature collection for model fitting.

```

# Subdivide presence/absence for training/testing
m.sf$kfold <- dismo::kfold(m.sf, k = 5)
pseudo.dismo$kfold <- dismo::kfold(pseudo.dismo, k = 5)

```

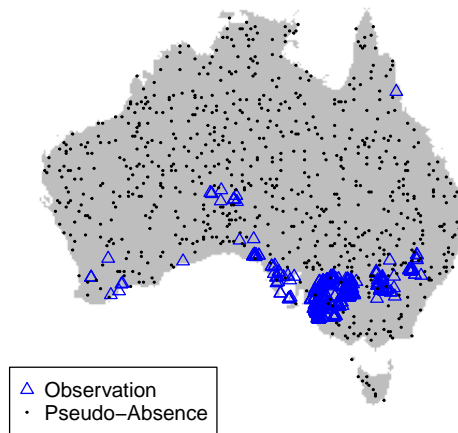



Figure 3: Observation and pseudo-absence point geo plot.

```
# Create presence/absence binomial dataset and add environmental data
present <- terra::subset(m.sf, select = "kfold")
present$pa <- 1
absent <- pseudo.dismo
absent$pa <- 0
names(absent) <- c("geometry", "kfold", "pa")
sf::st_geometry(absent) <- "geometry"
pa.data <- rbind(present, absent)
envt.data <- terra::extract(bioclim.h, pa.data, ID = FALSE)
pa.data <- cbind(pa.data, envt.data)
```

In total, ten coefficients were considered for modelling. The bestglm package was used to assess the most parsimonious model per Aikake Information Criteria. AIC balances a trade-off between the likelihood that a model could have produced the observations and the number of coefficients in the model (i.e., it seeks the most explanatory power for the least coefficients).

```
# Subset presence/absence data and run bestglm to select best model per AIC
fit <- as.data.frame(pa.data)
fit <- fit[, !(names(fit) %in% c("geometry", "kfold"))]
fit <- relocate(fit, pa, .after = last_col())
best.fit <- bestglm(fit, IC = "AIC", family = binomial(link = "logit"))
best.fit
```

```
## AIC
## BICq equivalent for q in (0.81076908163625, 0.972486680903518)
## Best Model:
##           Estimate Std. Error    z value    Pr(>|z|)
## (Intercept)  3.51550680 3.533624137   0.9948729 3.197981e-01
## bio1        -1.67998031 0.237603914  -7.0705077 1.543678e-12
## bio5         0.76953897 0.144002365   5.3439329 9.095127e-08
## bio6        -0.48702145 0.234198134  -2.0795275 3.756890e-02
```

```
## bio12      -0.01834490 0.004450777 -4.1217291 3.760393e-05
## bio13      0.10144880 0.018672596  5.4330316 5.540459e-08
## bio14      0.12476996 0.044384341  2.8111257 4.936849e-03
## elev      -0.01136999 0.001094301 -10.3901805 2.748325e-25
## sand       0.11965003 0.021891287  5.4656461 4.612236e-08
## ocd       -0.54089608 0.157349176 -3.4375527 5.869966e-04
```

```
colnames(best.fit$Subsets)[12] <- "logLike"
kableExtra::kable_styling(
  knitr::kable(
    best.fit$Subsets,
    align = "c",
    digits = 2,
    caption = "Logistic GLM AIC evaluation."),
  latex_options = c("scale_down", "hold_position"))
```

Table 2: Logistic GLM AIC evaluation.

	Intercept	bio1	bio3	bio5	bio6	bio12	bio13	bio14	elev	sand	ocd	logLike	AIC
0	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	-1013.38	2026.76
1	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	-615.95	1233.90
2	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	-475.23	954.47
3	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	-398.15	802.31
4	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	-365.32	738.63
5	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	-349.57	709.14
6	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	-344.18	700.37
7	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	-337.76	689.52
8	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	-330.65	677.30
9*	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	-328.46	674.92
10	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	-328.38	676.77

Model 9, which includes all variables except for Bio3, was selected since it had the lowest AIC score (Table 2). To validate, model 10 (another model with a low AIC score) was fit and compared to model 9.

```
# Fit models 9 and 10
m.9 <- glm(
  pa ~ bio1 + bio5 + bio6 + bio12 + bio13 + bio14 + elev + sand + ocd,
  data = pa.data,
  family = binomial(link = "logit"),
  subset = kfold != 1)

m.10 <- glm(
  pa ~ bio1 + bio3 + bio5 + bio6 + bio12 + bio13 + bio14 + elev + sand + ocd,
  data = pa.data,
  family = binomial(link = "logit"),
  subset = kfold != 1)
```

Model 10 had a slightly lower residual deviance (543.67 versus 543.9). But, the difference was minor and model 10 included an additional coefficient. AIC reasonably picked model 9 as the most parsimonious model.

Model 9 was fit and the results plotted.

```
# Fit the GLM
glm.model <- glm(
  pa ~ bio1 + bio5 + bio6 + bio12 + bio13 + bio14 + elev + sand + ocd,
```



```

data = pa.data,
family = binomial(link = "logit"),
subset = kfold != 1)

# Summarize and plot model results
summary(glm.model)

##
## Call:
## glm(formula = pa ~ bio1 + bio5 + bio6 + bio12 + bio13 + bio14 +
##      elev + sand + ocd, family = binomial(link = "logit"), data = pa.data,
##      subset = kfold != 1)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.830911   3.809106   1.006 0.314548
## bio1         -1.672875   0.267622  -6.251 4.08e-10 ***
## bio5          0.753690   0.159616   4.722 2.34e-06 ***
## bio6         -0.382116   0.255580  -1.495 0.134890
## bio12        -0.018041   0.004795  -3.762 0.000168 ***
## bio13         0.098314   0.019614   5.012 5.37e-07 ***
## bio14         0.130986   0.047808   2.740 0.006147 **
## elev         -0.011388   0.001218  -9.347 < 2e-16 ***
## sand          0.113872   0.023578   4.830 1.37e-06 ***
## ocd          -0.667088   0.179054  -3.726 0.000195 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 1566.5  on 1129  degrees of freedom
## Residual deviance:  543.9  on 1120  degrees of freedom
## (40 observations deleted due to missingness)
## AIC: 563.9
##
## Number of Fisher Scoring iterations: 8

par(mar = c(3, 3, 1, 1), mgp = c(2, 1, 0))
dismo::response(
  glm.model,
  fun = function(x, y, ...) predict(x, y, type = "response", ...))

```

The model results showed that all coefficients were statistically significant except for Bio6 (Min Temperature of Coldest Month). Examining response graphs (Figure 4), the direction of relationships mostly made biological sense.

As expected, probability of presence increased with Bio13 (Precipitation of Wettest Month), Bio14 (Precipitation of Driest Month), and sand soil composition. Probability of presence decreased with elevation (malleefowl habitat in low lying areas).

Other relationships were less intuitive. Contrary to expectations, probability of presence decreased with Bio12 (Annual Precipitation) and carbon density of soil. However, this may reflect a preference against damp forests rather than a preference against these variables within viable malleefowl habitat.

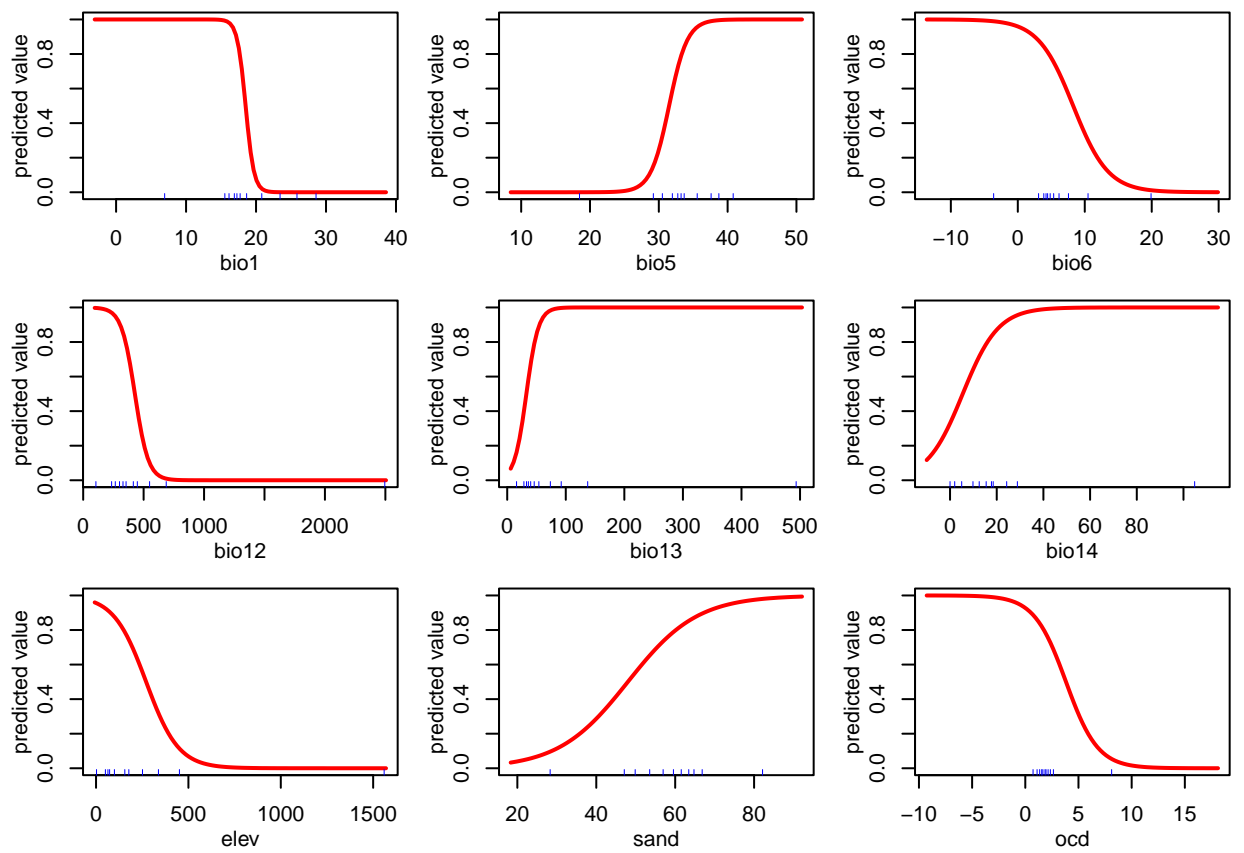


Figure 4: Logistic GLM results.

The most unexpected results were around temperature. Prior research suggested that increased winter temperature reduces breeding success (Stenhouse & Moseby, 2022). But, Bio6 was not statistically significant. Additionally, probability of presence increased with Bio5 (Max Temperature of Warmest Month) and decreased with Bio1 (Annual Mean Temperature). It's odd that the direction of effect reversed for two similar variables. More analysis is needed to explain this finding.

Next, the model was fit on historical data and an ROC curve was plotted (Figure 5).

```
# Create a prediction layer
glm.pred <- terra::predict(bioclim.h, glm.model, type = "response")
test.present <- sf::st_coordinates(subset(pa.data, pa == 1 & kfold == 1))
test.absent <- sf::st_coordinates(subset(pa.data, pa == 0 & kfold == 1))
glm.eval <- dismo::evaluate(
  p = test.present,
  a = test.absent,
  model = glm.model,
  x = bioclim.h)

# Plot ROC curve
plot(glm.eval, "ROC", type = "l")
```

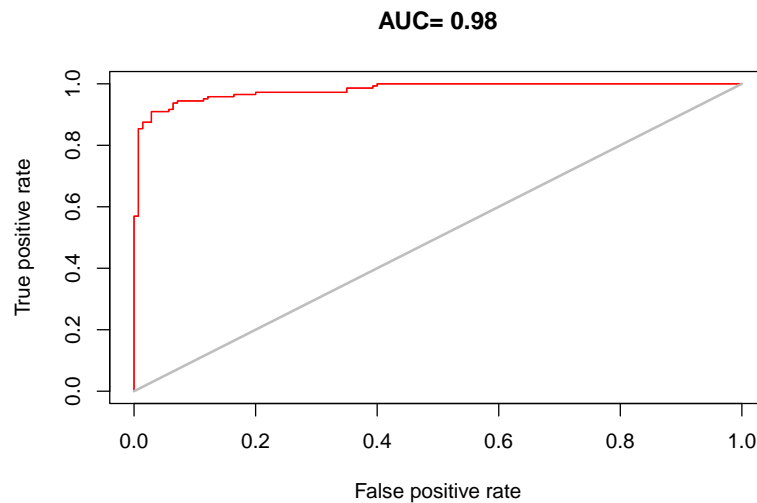


Figure 5: Logistic GLM ROC plot.

The area under the ROC curve was 0.98, indicating little trade-off between sensitivity and specificity across threshold of presence. Because there was so little trade-off, the threshold at which the sum of specificity and sensitivity were maximized was selected to determine presence in the final model.

```
# Extract threshold that maximizes specificity plus sensitivity
spec.sense <- plogis(threshold(glm.eval, stat = "spec_sens"))
```

This threshold was 0.664. Finally, estimated historical and future distributions were plotted (Figure 6).

```
# Create maps based on threshold that maximizes specificity plus sensitivity
glm.map <- glm.pred >= spec.sense
glm.pred.future1 <- predict(bioclim.f1, glm.model, type = "response")
```

```

glm.map.future1 <- glm.pred.future1 >= spec.sense
glm.pred.future2 <- predict(bioclim.f2, glm.model, type = "response")
glm.map.future2 <- glm.pred.future2 >= spec.sense

# Plot maps
par(mfrow = c(2, 2), mar = c(.5, .5, .5, .5))
terra::plot(
  glm.map,
  legend = FALSE,
  col = c("grey", "blue"),
  axes = FALSE,
  main = "1970-2000 Distribution",
  font.main = 1,
  cex.main = 1)
add_legend(
  "bottomleft",
  legend = "Est. Malleefowl Presence",
  fill = "blue",
  cex = .7)
terra::plot(
  glm.map.future1,
  legend = FALSE,
  col = c("grey", "blue"),
  axes = FALSE,
  main = "2021-2040 Distribution",
  font.main = 1,
  cex.main = 1)
add_legend(
  "bottomleft",
  legend = "Est. Malleefowl Presence",
  fill = "blue",
  cex = .7)
terra::plot(
  glm.map.future2,
  legend = FALSE,
  col = c("grey", "blue"),
  axes = FALSE,
  main = "2041-2060 Distribution",
  font.main = 1,
  cex.main = 1)
add_legend(
  "bottomleft",
  legend = "Est. Malleefowl Presence",
  fill = "blue",
  cex = .7)

# Create masks
f1.msk <- ifel(glm.pred.future1 < spec.sense, NA, 1)
f1m <- terra::mask(glm.map.future1, f1.msk)
glm.map.future1.p <- terra::mask(f1m, f1m)
f2.msk <- ifel(glm.pred.future2 < spec.sense, NA, 1)
f2m <- terra::mask(glm.map.future2, f2.msk)
glm.map.future2.p <- terra::mask(f2m, f2m)

```

```
# Plot overlaid scenario map
terra::plot(
  glm.map,
  legend = FALSE,
  col = c("grey", "blue"),
  axes = FALSE,
  main = "Time Period Distribution Overlay",
  font.main = 1,
  cex.main = 1)
terra::plot(glm.map.future1.p, legend = FALSE, add = TRUE, col = "green")
terra::plot(glm.map.future2.p, legend = FALSE, add = TRUE, col = c("yellow"))
add_legend(
  "topleft",
  legend = c("1970-2000", "2021-2040", "2041-2060"),
  fill = c("blue", "green", "yellow"),
  cex = .8,
  bg = "white")
```

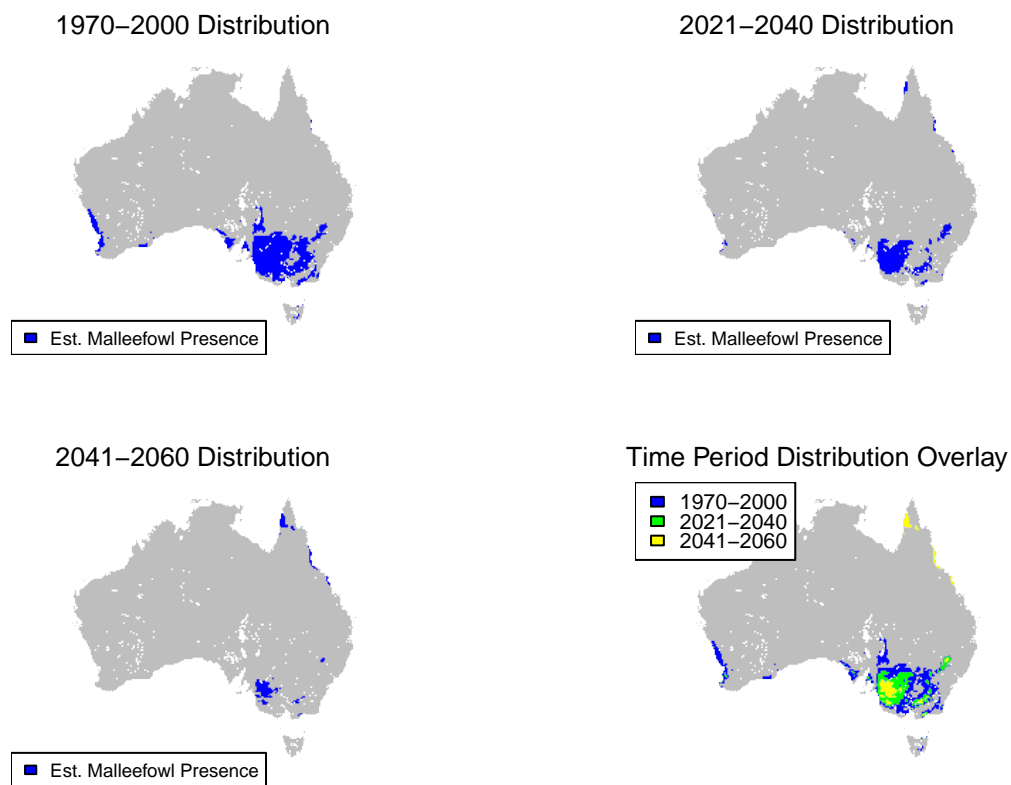


Figure 6: Malleefowl distribution plots.

Results suggested that malleefowl distribution will contract significantly in both 2021-2040 and 2040-2060.

```
# Table of shift from 1970-2000 to 2021-2040 and from 1970-2000 to 2041-2060
h.f1 <- arrange(
```

```

data.frame(
  table(values(glm.map), values(glm.map.future1))),
Var1,
Var2)
colnames(h.f1) <- c("Hist", "F1", "Freq")
h.f2 <- arrange(
  data.frame(
    table(values(glm.map), values(glm.map.future2))),
Var1,
Var2)
colnames(h.f2) <- c("Hist", "F2", "Freq")

```

Specifically, malleefowl habitat will decline to 46% of its historical range by 2021-2040 and to 17% by 2041-2060.

In particular, Western Australia is forecast to be uninhabitable by malleefowl by 2040-2060. The largest viable habitat is expected to be in western New South Wales. However, some new viable habitat is expected to develop in coastal Queensland.

Statistical Analysis of Megapodiidae Extinction Risk Versus Dispersal Ability

The Hand-Wing Index is a commonly used proxy for avian dispersal ability, which is thought to influence extinction risk (i.e., more mobile species are at less risk of extinction) (Sheard et al., 2020).

However, this relationship may not hold among Megapodiidae species. While some are strong flyers, many are heavy and poor at flying (Harris, Burks & Leache, 2014). If wing morphology does not proxy well for dispersal ability due to weight, HWI may be a poor predictor of extinction risk.

To evaluate, data on Evolutionary Distinct Globally Endangered (EDGE) scores, which are a measure of species' evolutionary uniqueness and extinction risk, and HWI were analyzed. Data for 30 Megapodiidae individuals across 10 species were provided by Imperial College.

First, data were consolidated and average HWI per species was taken (Table 3).

```

# Pull in data, filter for Megapodiidae, and arrange by EDGE score
setwd("C:/Users/Michael Jordan/Desktop/R/BioCompAssessment")
b.full <- read.csv("birds_edge_trait.csv", header = TRUE)
b <- b.full[b.full$Family == "MEGAPODIIDAE", c(2,5:7)]

# Aggregate by species and average across trait measurements
b.agg <- dplyr::arrange(
  dplyr::summarise(
    group_by(b, Species, Common.name, EDGE),
    .groups = "keep",
    avg.HWI = mean(Hand.wing.Index)),
  desc(EDGE))
names(b.agg)[2] <- "Common Name"

# Output results
kableExtra::kable_styling(
  kableExtra::row_spec(
    knitr::kable(
      b.agg,
      digits = c(2, 1),

```



```

align = "c",
caption = "Megapodiidae species EDGE score and average HWI."),
0,
bold = TRUE),
position = "center",
latex_options = c("hold_position"))

```

Table 3: Megapodiidae species EDGE score and average HWI.

Species	Common Name	EDGE	avg.HWI
Macrocephalon maleo	Maleo	5.59	22.0
Leipoa ocellata	Malleefowl	4.80	31.0
Aepyodius bruijnii	Waigeo Brush-turkey	4.76	19.5
Eulipoa wallacei	Moluccan Scrubfowl	4.56	33.5
Megapodius laperouse	Micronesian Scrubfowl	4.50	31.4
Megapodius pritchardii	Tongan Scrubfowl	3.81	18.8
Megapodius layardi	Vanuatu Scrubfowl	3.81	25.3
Megapodius bernsteinii	Sula Scrubfowl	3.81	26.0
Megapodius geelvinkianus	Biak Scrubfowl	3.80	18.6
Megapodius nicobariensis	Nicobar Scrubfowl	3.79	20.1

No relationship between EDGE and HWI was evident. To evaluate, a linear model of EDGE score on HWI was fit.

```

# Fit and summarise linear model
m.b <- lm(EDGE ~ avg.HWI, data = b.agg)
summary(m.b)

##
## Call:
## lm(formula = EDGE ~ avg.HWI, data = b.agg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.5522 -0.3905 -0.1841  0.2190  1.3387
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.62100     0.93541   3.871  0.00473 **
## avg.HWI      0.02857     0.03711   0.770  0.46352
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6343 on 8 degrees of freedom
## Multiple R-squared:  0.06897,    Adjusted R-squared:  -0.04741
## F-statistic: 0.5927 on 1 and 8 DF,  p-value: 0.4635

```

As expected, HWI was not a statistically significant predictor of EDGE score. The p-value of HWI was 0.464, above an alpha of 5%, causing the null hypothesis of no relationship between HWI and EDGE score to fail to be rejected.

Additionally, the model itself was a poor fit for the data. R-squared was 0.069, demonstrating that the model explained only a small percentage of variation. This is evident in plotting (Figure 7).

```
# Create a scatter plot of the data and add the linear regression
par(mfrow = c(1, 1), mar = c(4, 4, 0, 0))
plot(b.agg$avg.HWI, b.agg$EDGE, xlab = "avg.HWI", ylab = "EDGE")
abline(m.b)
```

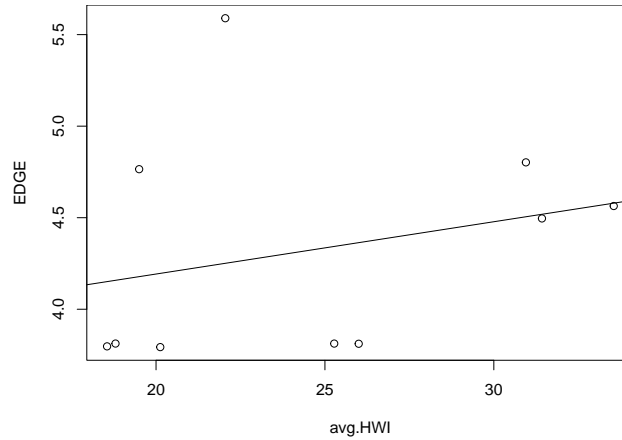


Figure 7: Scatter plot with linear regression of EDGE score on HWI.

To validate the applicability of linear regression as a technique to these data, model diagnostics were plotted and assessed (Figure 8).

```
# Plot diagnostics
par(mfrow = c(1,2), mar = c(4, 4, 0, 2))
layout(
  matrix(c(1, 2), nrow = 1),
  widths = c(1, 1),
  heights = c(1, 1),
  TRUE)
plot(m.b, which = 1:2)
```

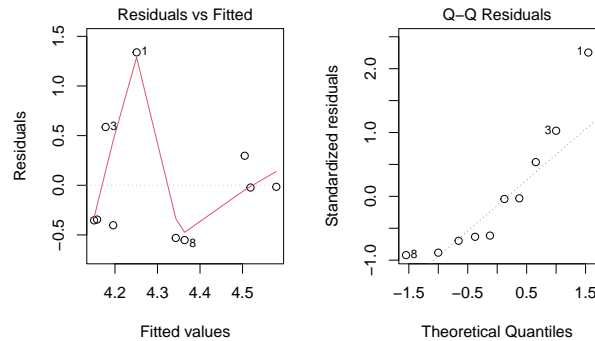


Figure 8: Diagnostics of linear regression of EDGE score on HWI.

This demonstrated two issues.

First, the plot of residuals versus fitted values exhibited heteroskedasticity. Variance is much higher on the low end of the distribution than on the high end. Second, the Q-Q plot of z-transformed residuals versus

theoretical quantiles of the z distribution showed that the residuals are not normally distributed. If residuals were normally distributed, they would line up along the straight line. Instead, they curve off it at extremes of the distribution.

These issues will cause the estimate of model standard error (and therefore p-values and confidence intervals) to be biased. As a result, the model cannot be used to evaluate the statistical significance of HWI on EDGE score.

Summary

Malleefowl are currently listed as a vulnerable species per the IUCN Red List and Australian law. Although overall population size has not been fully determined (Benshemesh, 2007), best evidence suggests it has declined significantly from its historic range and is continuing to decline. The Australian government has instituted a recovery plan to increase malleefowl population.

This species distribution analysis suggests three potential adjustments to the recovery plan. First, because Western Australia is expected to become nonviable to malleefowl by 2040, conservation efforts there should be abandoned if there are funding shortfalls. Second, because New South Wales is expected to contain the largest remaining malleefowl habitat by 2060, conservation efforts should be concentrated there. Third, because some parts of Queensland are expected to become habitable by malleefowl by 2060, the plan should consider re-introduction efforts there in the future.

Finally, this report assessed whether Megapodiidae extinction risk was related to dispersal ability. No relationship was expected due to many Megapodiidae species being heavy and poor at flying (Harris, Burks & Leache, 2014). Of course, a negative cannot be proven. But, a linear regression failed to demonstrate any statistically significant effect of dispersal ability on Megapodiidae extinction risk.

Reference List

- Benshemesh, J. (2007) National Recovery Plan for Malleefowl. *Department for Environment and Heritage, South Australia*. <https://www.dcceew.gov.au/sites/default/files/documents/malleefowl.pdf>
- Booth, D.T. (1986) Effect of Temperature on Development of Mallee Fowl *Leipoa ocellata* Eggs. *Physiological Zoology*. 60 (4), 437-445. <https://www.jstor.org/stable/30157905>
- GBIF (2023) Global Biodiversity Information Facility. <https://www.gbif.org/> [Accessed: November 28, 2023].
- Harris, R.B., Birks, S. & Leaché, A.D (2014) Incubator birds: biogeographical origins and evolution of underground nesting in megapodes (Galliformes: Megapodiidae). *Journal of Biogeography*. 41, 2045-2056. <https://doi.org/10.1111/jbi.12357>
- Parsons, B.C., Short, J.C. & Roberts, J.D. (2008) Contraction in the range of Malleefowl (*Leipoa ocellata*) in Western Australia: a comparative assessment using presence-only and presence—absence datasets. *Emu - Austral Ornithology*. 108 (3), 221-231. <https://doi.org/10.1071/MU08002>
- Sheard, C., Neate-Clegg, M.H.C., Aloravainen, N., Jones, S.E.I.J., Vincent, C., MacGregor, H.E.A., Bregman, T.P., Claramunt, S. & Tobias, J.A. (2020) Ecological drivers of global gradients in avian dispersal inferred from wing morphology. *Nature Communications*. 11, 2463. <https://doi.org/10.1038/s41467-020-16313-6>
- Stenhouse P. & Moseby K. (2022) Trends in breeding activity of the threatened Malleefowl (*Leipoa ocellata*): what can we expect under a changing climate? *Emu - Austral Ornithology*. 122 (1), 51-60. <https://doi.org/10.1080/01584197.2022.2045870>