Code

How do IUCN Range maps of birds compare to SDMs in Indonesian New Guinea? What are the main predictors for this distribution?

Jonathan Gehret, Matr.-Nr.: 23301512, Paul Leister

31.08.2021

1. Introduction

1.1 Study area: West New Guinea (Papua & Papua barat)

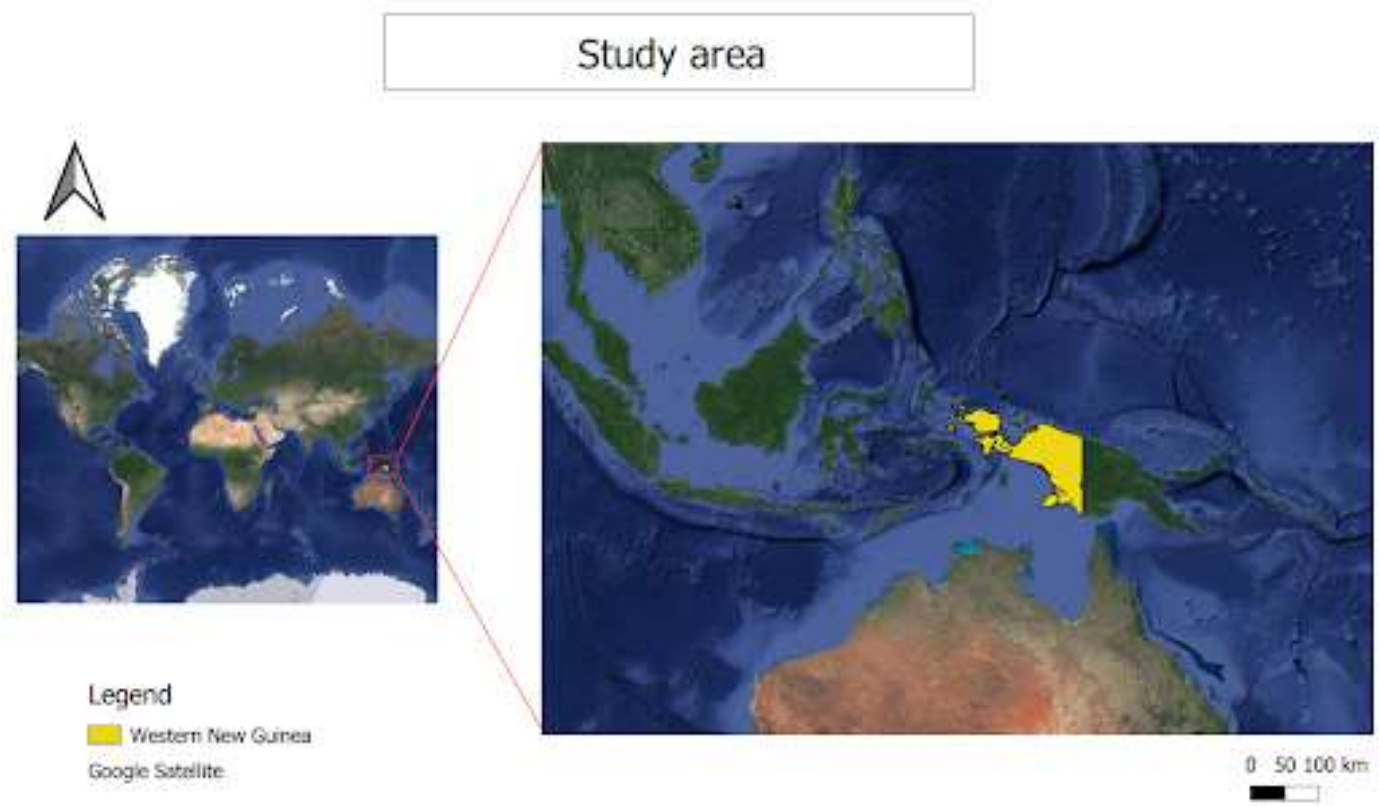


Fig. 1. The study area of Western New Guinea

Papua & Papua barat

Papua and Papua Barat are both provinces of Indonesia on the western part of the island of new guinea. The island is the second largest in the world and is geographically counted as part of Australia. An extremely high biodiversity can be found here. This is due to the intact rainforests, the mangrove swamps in the south and the coral reefs. The island was formed due to plate tectonic movement (pacific plate under the Indian-Australian plate). Therefore, the island is crossed by the Maoke Mountains and has a height of 4884 m with the Puncak Jaya. The highest population density is reached on the coasts, the mountainous areas are rather sparsely populated. http://www.papuaweb.org/ (http://www.papuaweb.org/)

In 1660 they were part of the dutch new guinea colony and got concurred and occupied by the dutch east India company. This colony got attacked by the Indonesian military in 1961, with the purpose of the reunition of the corelated tribes. With the mediation of the united nations and a referendum of the residents (act of choice) Papua and papua barat became part of Indonesia. https://pm20.zbw.eu/category/geo/i/141619/ (https://pm20.zbw.eu/category/geo/i/141619/) https://gis.dukcapil.kemendagri.go.id/peta/ (https://gis.dukcapil.kemendagri.go.id/peta/)

Papua Papua barat Size: 319.036,05 km² Size: 115.364 km² Citizen: 4.354.468 Citizen: 1.149.282 Citizenrange: 14 Citizen/km² Citizenrange: 10 Citizen/km² Capital: Jayapur Capital: Manokwari https://gis.dukcapil.kemendagri.go.id/peta/ (https://gis.dukcapil.kemendagri.go.id/peta/)

2. Theory

2. Methods

2.x Creation process?

Gibts sowas zu schreiben? Zum finden der ARten: loops for plotting all species (iusn sowie GBIF) Erster Ansatz mit Kasuaren, aber zu wenig Daten! -> in the case of low amount of observation data, some research has been done on how to proceed in cases like this. While this is definitely a very interesting area of study, we decided that pursuing these approaches would exceed the time-investment deemed reasonable to invest into this work. It could be pursued in future work. (Briscoe, Fois, Lomba) Entscheidung, Pseudo-absenzdaten selbst zu basteln anstatt biomod2 dafür zu verwenden First the scripts for the singular steps were created, then for continuous use going through the markdown file some of the scripts were changed to be functions and the markdown script was built in a way to easily create those SDMs with our methods for any chosen bird species of the GBIF data set! Only slight modifications are needed: Change of species names, adjustment of some variable names, names of images to be plotted Used libraries: Biomod2 for SDM creation with plenty of nice relevant tools Raster for work with the raster data Sf for work with spatial points Rgdal for reading in and working with shapefiles, eg readOGR What about rgeos, spatialeco, dplyr -> do we need these Workaround for better image alignment was saving some images in a folder and the plotting them. Could also plot directly from loop, as was done in some other cases, however alignment wasn't so nice then.

2.z : .rmd: markdown desciritio

All the scripts were put into one markdown file, namely MCMMB.rmd. Visual appeal was maximised by setting of plot sizes etc. Use of cache, fig.width, fig.height. e.g.. Hiding of warnings, results.

Important code was displayed in the .rmd, other code can be found in the appendix.

Careful consideration was taking on which plots to include and which to leave out/put into the appendix.

For the creation of this HTML document, the MCMMB.Rmd was knitted.

To find the species to be used for our analysis we used the loops in *bird_plot_loops.R* to plot all species from the Gbif and IUCN data sets to find birds with sufficient observation points and different ranges to be used for our project.

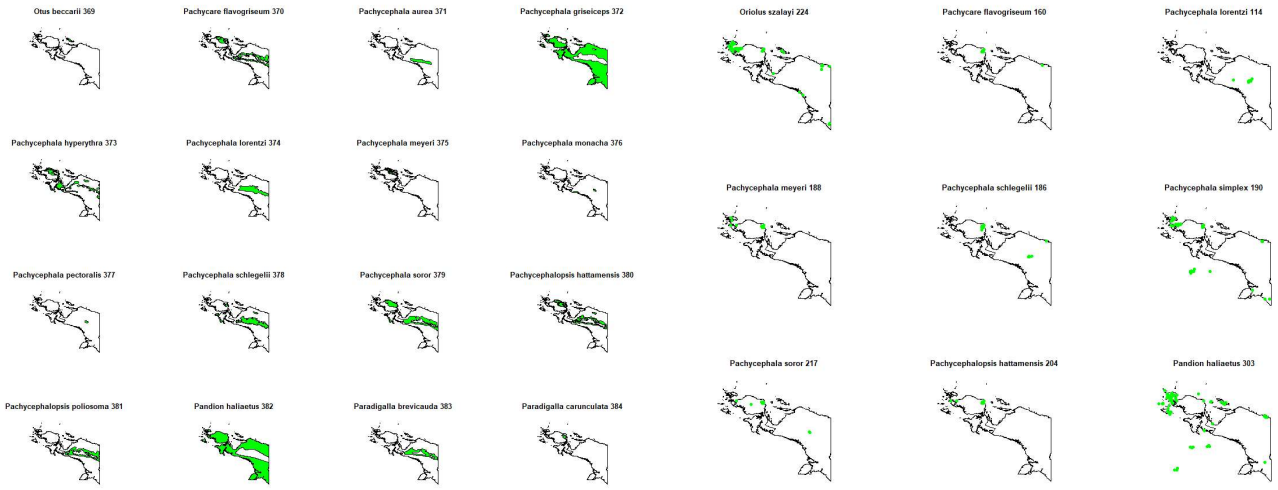


Fig. X. Excerpt of the plotted maps for finding suitable species, containing the Pachycephala Genus; IUCN (left), GBIF (right)

2.1. Main script

The required libraries and the setup of the IUCN and GBIF data sets are done in the *MCMMB_Main.R* script. For the IUCN data, the dataset was provided beforehand. It is read in here and named *birdlife*. Additionally a dataframe including all 555 species names that appear in the birdlife dataset was created. A shapefile containing the extent of the observation area is read in and named *regio*.

Hide

```
source("MCMMB_Main.R")
```

The GBIF data were downloaded manually beforehand from the website, however the r library *rgbif* could be used instead. After cleaning up the GBIF data by only retaining the important columns of gbifID, species names and coordinates using *select()* and removing all data points without coordinates, spatial points are created from the data. These are then cropped to the observation area by use of *regio*. Of these, only species with more than 100 occurrences are kept. The names of those species are added to the *species* object.

Hide

```
# example code from MCMMB_Main.R for filtering out all species with less than 100 occurences
gbif_crop = gbif_crop_all %>% group_by(species) %>% filter(n() >= 100 ) %>% ungroup()
```

2.2. Species data

The script species_loops.R includes two functions each for creating lists with the input bird species data for GBIF (*get_gbif_birds(bird_names,gbif_crop,regio*) and IUCN (*get_iucn_birds(bird_names,birdlife,regio*) and plotting them, respectively. IUCN birdlife data is depicted in red, GBIF points in green. The *gbif_birds* list contains the GBIF observation points for the selected species, which are then used for the creation of pseudo-absence data and further for the creation of the SDMs. The resulting *iucn_birds* list is only used again for the final comparison of calculated SDMs with IUCN birdlife data.

Hide

```
# loading of the described functions
source("species_loops.R")

# cassowaries didn't have enough data
#bird_names = c("Casuarius bennetti", "Casuarius casuarius", "Casuarius unappendiculatus")

# we chose Pachycephala instead
bird_names = c("Pachycephala lorentzi",
               "Pachycephala meyeri",
               "Pachycephala schlegelii",
               #"Pachycephala simplex", # this species doesn't appear in the birdlife data se
t on the island
               "Pachycephala soror")

# these names can be changed to any other bird species occuring in the data sets, for fully a
utomated creation of sdms following this markdown file:
# bird_names = c(your_favorite_species_here)

# for plotting GBIF data next to IUCN data
par(mfrow = c(2,length(bird_names)))

# create list with birds out of the IUCN and GBIF data and plot them
iucn_birds = get_iucn_birds(bird_names,birdlife,regio)
gbif_birds = get_gbif_birds(bird_names,gbif_crop,regio)
```

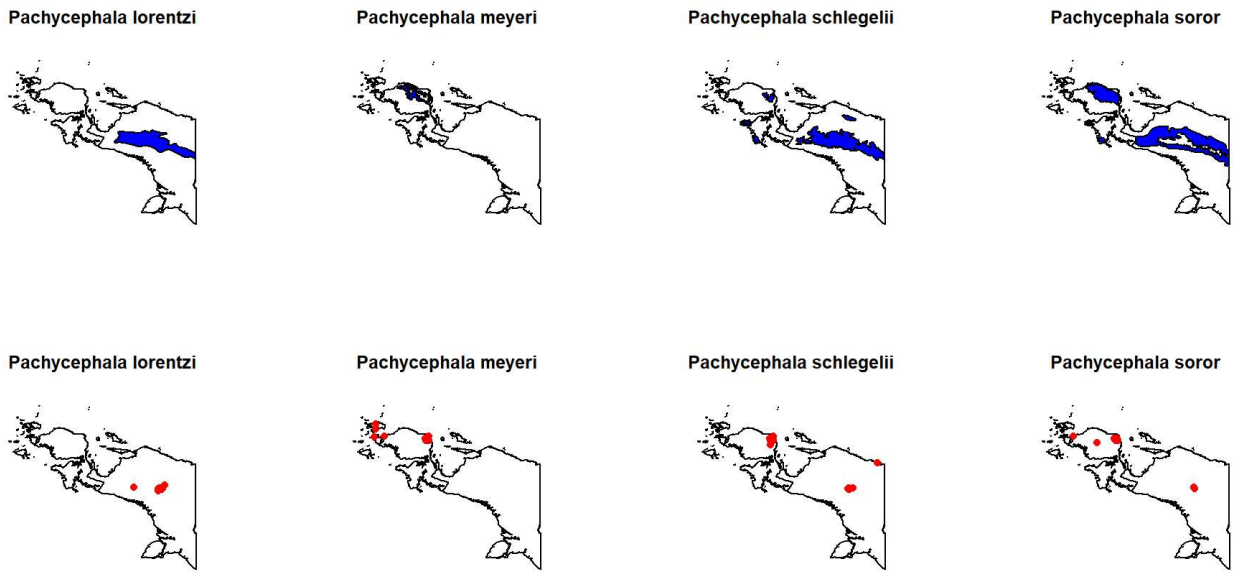


Fig. X. The IUCN birdlife distribution ranges for the four selected species in blue. The GBIF observation data in red.

2.3. Create presence-absence data

While *biomod2* also brings with it capabilities to create pseudo-absence data, we built a script to do it, following the tutorial of the MCMMB lecture. First a raster grid of the region was needed, which was created by masking the elevation layer (compare 2.4.) over the regio shapefile of the observation area. From this mask centroids in the WGS84 projection were created. Based on these centroids the newly evoked *presence_absence_list* is populated by use of a for loop. As the creation of the pseudo-absence data is chance-based, a seed was set. For every species included in the *gbif_birds* list handed over to *create_pseudo_absence* function, the pseudo-absence data is created at the randomly located centroids. The number of absence points is set equal to the number of presence points. These data points are then plotted with the real presence data, colored in green and red respectively. As a result the *presence_absence_list* is returned.

Hide

```
# example from presence_absence.R showcasing the creation of pseudo absences by use of the ea
rlier created centroids (centroids_all) and the occurence data mapped to the centroids (centr
oids_occ) and combining both
presence_absence = rbind(centroids_all[sample(nrow(centroids_all),nrow(centroids_occ), replac
e = TRUE),],
                         centroids_occ)
```

```
"Be sure to take care when considering the use of pseudo-absences versus true absences for sp
ecies distribution
modeling. Similarly, it is extremely important to consider the influence of sampling bias in
the data used to
train models. Further reading: e.g., Guillera-Arroita et al. 2015, Kramer-Schadt et al. 2013,
and Merow et al
2013."
```

Hide

```
# load presence_absence.R script including the function
source("presence_absence.R")
par(mfrow = c(1,length(bird_names)))
presence_absence_list = create_pseudo_absence(regio,gbif_birds)
```

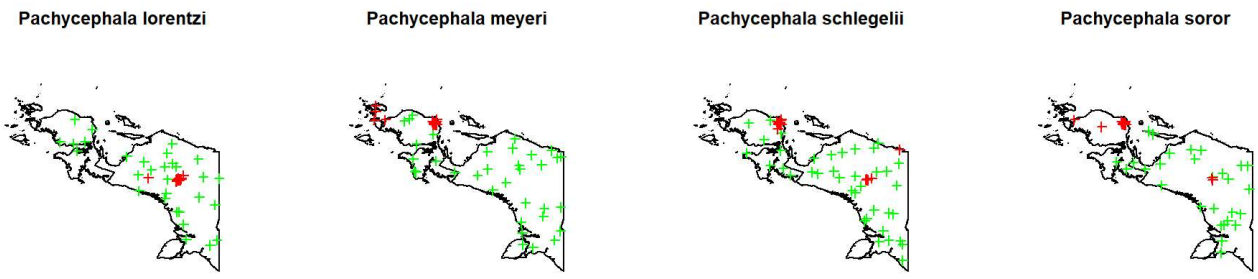


Fig. X. The centroids of the created pseudo-absence data (green) and the GBIF presence data (red) for the selected species.

2.4. Indicators

The *Indicator.R* script is used to read in data for the various indicators (precipitation, elevation, landcover, primary_forest, temperature, human population) to be used for the SDM. These are then cropped to the target region and saved in *.tif* format in the folder *data/indicator_stack/* for later use. As the process for some of the bigger files takes a lot of time, ready to use indicator *.tif* files are to be found in the aforementioned folder. These will then be loaded and stacked later on before the SDM creation.

Hide

```
# where all the predictors are loaded in
source("Indicator.R")
```

Hide

```
# stacking all predictors found in folder indicator stack in .tif format
# predictors: elevation, precipitation, temperature, primary forest, landcover, population
tif_predictors = stack(list.files(path = "data/indicator_stack/",
                                  full.names = TRUE,
                                  pattern = ".tif"))
plot(tif_predictors)
```

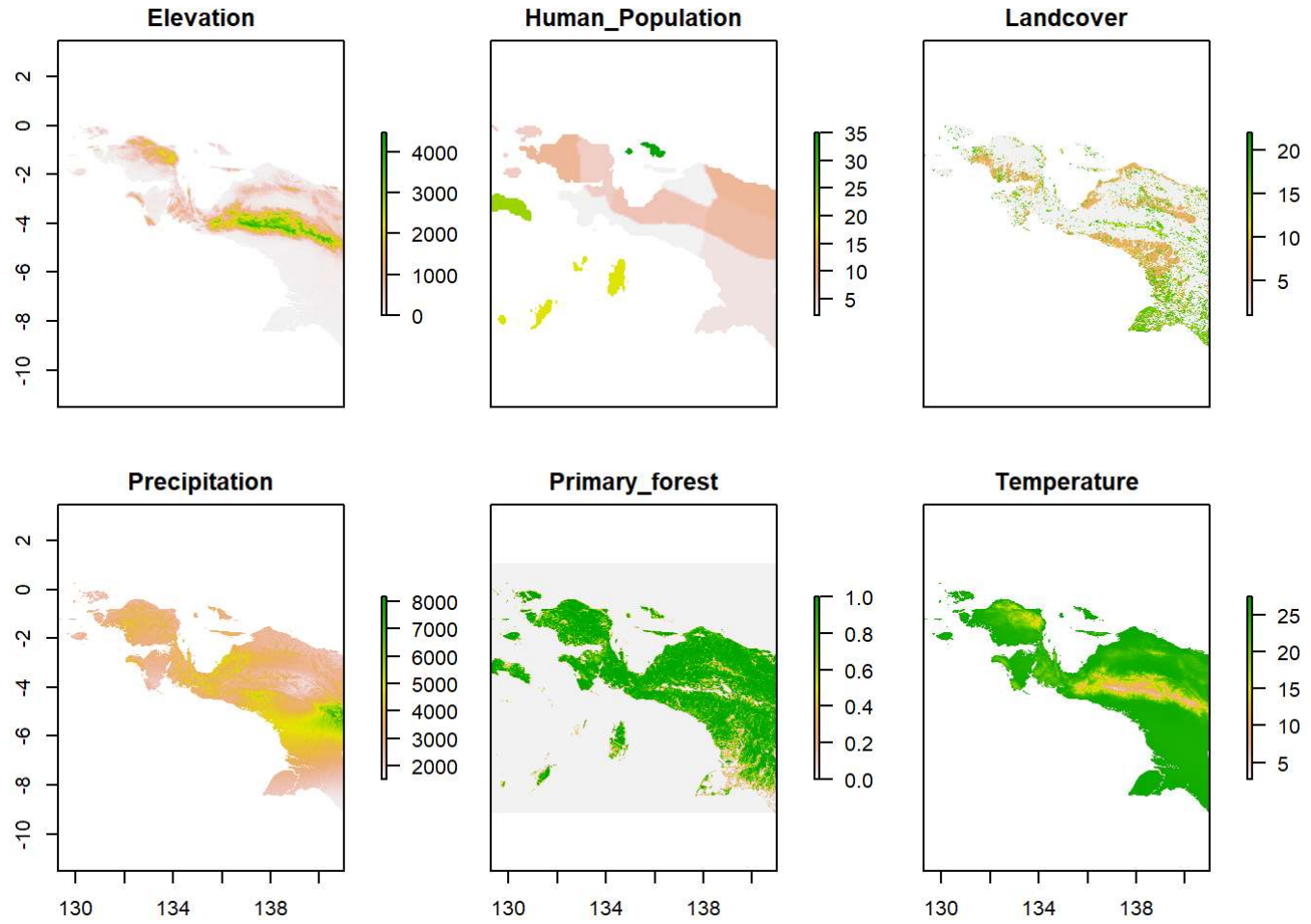


Fig. X. The used predictor data: Elevation with height in m, Human Population in XXXX??, land cover in over 20 different categories (noticeably XXXX), Precipitation in XXXX, Primary forest density and temperature in °C.

2.5. Species distribution models

The following was done closely following the script by Griffith (2017).

The script *sdm_biomod2.R* contains a function *calculate_sdm* which will calculate and return SDMs for any given presence absence data made of spatial points and stacked predictors in the *.tif* format. Various helper lists are created for storing the various outputs calculated in the function. There's also the option to change the settings for any of the used SDM methods, i.e. the GAM *k* value was changed to four, as the default one created an error message and four is a generally used value. Taking these input data, the function then iterates over every species handed to it by the following steps: First the input data is formatted by combining it into one object *format_bm*.

Hide

```
# loading BIOMOD script and passing models to test_sdms
source("sdm_biomod2.R")
gbif_sdms = calculate_sdm(presence_absence_list, tif_predictors)
```

These combined data are then called by *BIOMOD_Modeling()*, where the desired models can be set, as well as the earlier decided upon BIOMOD options are taken in. The option for splitting off validation data can also be set here, as well as the option to save the models. In this work the models GLM, GAM, ANN and RF were used. Alternative available methods can be shown using *BIOMOD_ModelingOptions()*.

Hide

```
# not to be executed example code taken from sdm_biomod2.R for creation of the SDMs
biomodels_1 = BIOMOD_Modeling(data = format_bm,
                              models = c('GLM','GAM','ANN','RF'),
                              SaveObj = TRUE,
                              models.options = myBiomodOptions,
                              # DataSplit = 80, # could be used for validation data
                              VarImport = 1)
```

Once these models have been calculated and put into *biomodels_1*, some further operations are available: The variable importance can be calculated (compare 3.2.), the models can be put together into an ensemble model using *BIOMOD_EnsembleModeling()*. This is done with a variable threshold of 0.6 for the predictors. For visualization purposes, the SDMs can also be projected by use of *BIOMOD_Projection()*, taking into account again the predictors of the observation area. Alternatively other areas with their respective predictors could be used to test the models on. The models, the variance importance, the ensemble models and the projections are returned by the function in the *output_list*.

3. Results

3.1. Evaluations

```
"We will focus TSS (see Allouche et al. 2006 for a comparison of all three). TSS is the sum o
f the rates that
we correctly classified presences and absences, minus 1. Higher is better (in the range -1 to
1), and represents
a balance between model maximizing sensitivity and specificity." Griffith (2017)

# add to markdown?? use other indicators??
#biomod_eval = get_evaluations(biomodels_1)
#biomod_eval["TSS","Testing.data",,,]
#biomod_eval["KAPPA","Testing.data",,,]
#biomod_eval["ROC","Testing.data",,,]
#evaluate(biomodels_1, data, stat, as.array=FALSE)

" GLM   GAM   ANN    RF
0.725    NA 0.757 0.938 "
#>     biomod_eval["TSS","Testing.data",,,]
#GLM   GAM   ANN    RF
#1.000 1.000 1.000 0.967
#>     biomod_eval["KAPPA","Testing.data",,,]
#GLM   GAM   ANN    RF
#1.000 1.000 1.000 0.967
#>     biomod_eval["ROC","Testing.data",,,]
#GLM   GAM   ANN    RF
#1.000 1.000 1.000 0.999

    #get_evaluations(biomod_ensemble)
"$Test.Spec_EMmeanByTSS_mergedAlgo_mergedRun_mergedData
      Testing.data Cutoff Sensitivity Specificity
KAPPA        0.846    564       93.75      90.909
TSS          0.847    564       93.75      90.909
ROC          0.979    562       93.75      90.909
"
```

```
## , , GLM, Full, AllData
##
##       Testing.data Cutoff Sensitivity Specificity
## KAPPA            1    495         100         100
## TSS              1    495         100         100
## ROC              1    500         100         100
##
## , , GAM, Full, AllData
##
##       Testing.data Cutoff Sensitivity Specificity
## KAPPA            1    495         100         100
## TSS              1    495         100         100
## ROC              1    500         100         100
##
## , , ANN, Full, AllData
##
##       Testing.data Cutoff Sensitivity Specificity
## KAPPA            1    495         100         100
## TSS              1    495         100         100
## ROC              1    500         100         100
##
## , , RF, Full, AllData
##
##       Testing.data Cutoff Sensitivity Specificity
## KAPPA            1    515         100         100
## TSS              1    515         100         100
## ROC              1    515         100         100
```

```
## $Pachycephala.lorentzi_EMmeanByTSS_mergedAlgo_mergedRun_mergedData
##       Testing.data Cutoff Sensitivity Specificity
## KAPPA            1  505.0         100         100
## TSS              1  505.0         100         100
## ROC              1  503.5         100         100
```

Table 1 Evaluation values KAPPA, TSS and ROC for select SDMs and ensemble SDMs.

3.2. Variance importance

```
"We can also calculate variable importances to compare influences of individual predictor var
iables within and
among models. For a publication, you might also create partial dependence plots. Do the diffe
rent models
agree on the importance of the variables?" Griffith (2017)
```

The bar plots displaying the variance importance show remarkable variations between different species and different models. Depending on the model used, the results vary considerably: While elevation in almost all cases appears to be the most important predictor, other predictors like temperature and rainfall can also have significant impact on the modelled distribution range. Land cover and primary forest generally didn't have a lot of impact. Noticeable is also the difference between some approaches in weighting the predictors, as temperature is weighted more highly by GLM in some cases, while the other approaches prefer elevation. This consequently shows in the depiction of the computed SDMs (see 3.3.).

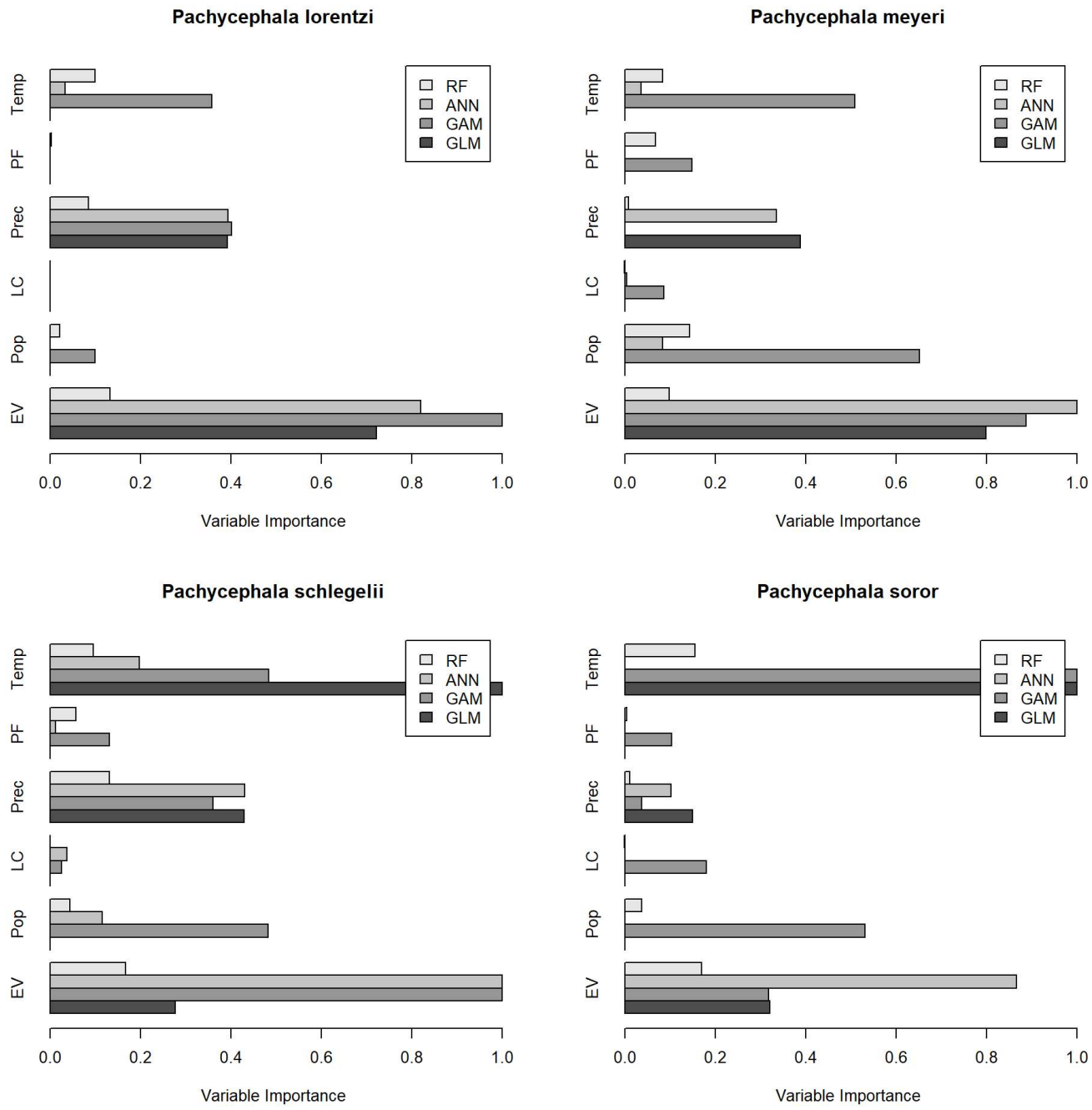


Fig. X. Variable importances for the different model approaches and species. Temperature (Temp), primary forest (PF), precipitation (Prec), land cover (LC), population density (Pop) and Elevation (EV) are decpicted.

3.3. SDM plots

```
"In order to visualize the model outputs, we should now use the SDMs to predict species occur
rence across
space using our modern environmental data. The models predict probabilities of occurrence, an
d it is
important to remember the interpretation of these outputs depend on the specific model, our d
ata, and our
assumptions (e.g., see discussion of Maxent outputs in Merow et al 2013)." Griffith (2017)
```

The plots of the SDMs show what the variable importances already suggest: the different models have at times greatly varying discrepancies in their predicted distribution areas. The legend indicates high confidence with higher values, dark green being the most sure the model can be and light beige being less confident. While the

different models mostly agree on P. lorentzi appearing almost exclusively in the central mountain ranges, the other species have a more varying predicted range, reaching up to Papua Barat. The marked areas especially in northern Papua Barat and the central mountain ranges are quite striking.

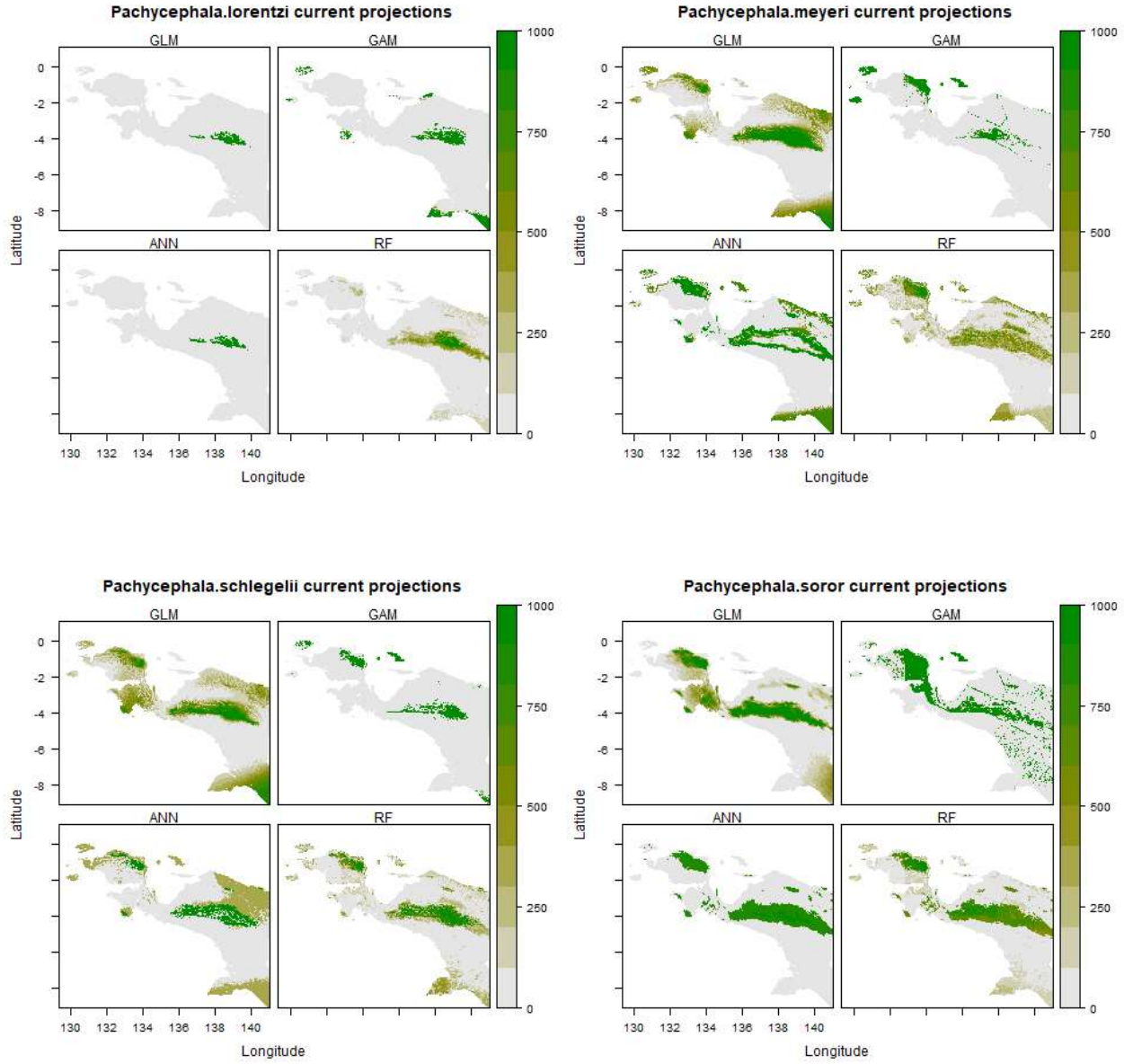


Fig. X. Plots of the species with subplots for the different SDMs; Greener color means higher confidence that the birds occur in that area.

3.4. Ensemble plots

```
"One approach for using the information in these various models is to combine them into an en
semble, or
collection of models merged together (Thuiller et al. 2009). We can take all models above a g
iven “quality”
threshold and combine them." Griffith (2017)
```

The Ensemble plots are put together from the other SDMs above a threshold of 0.6 in the variable importance section.

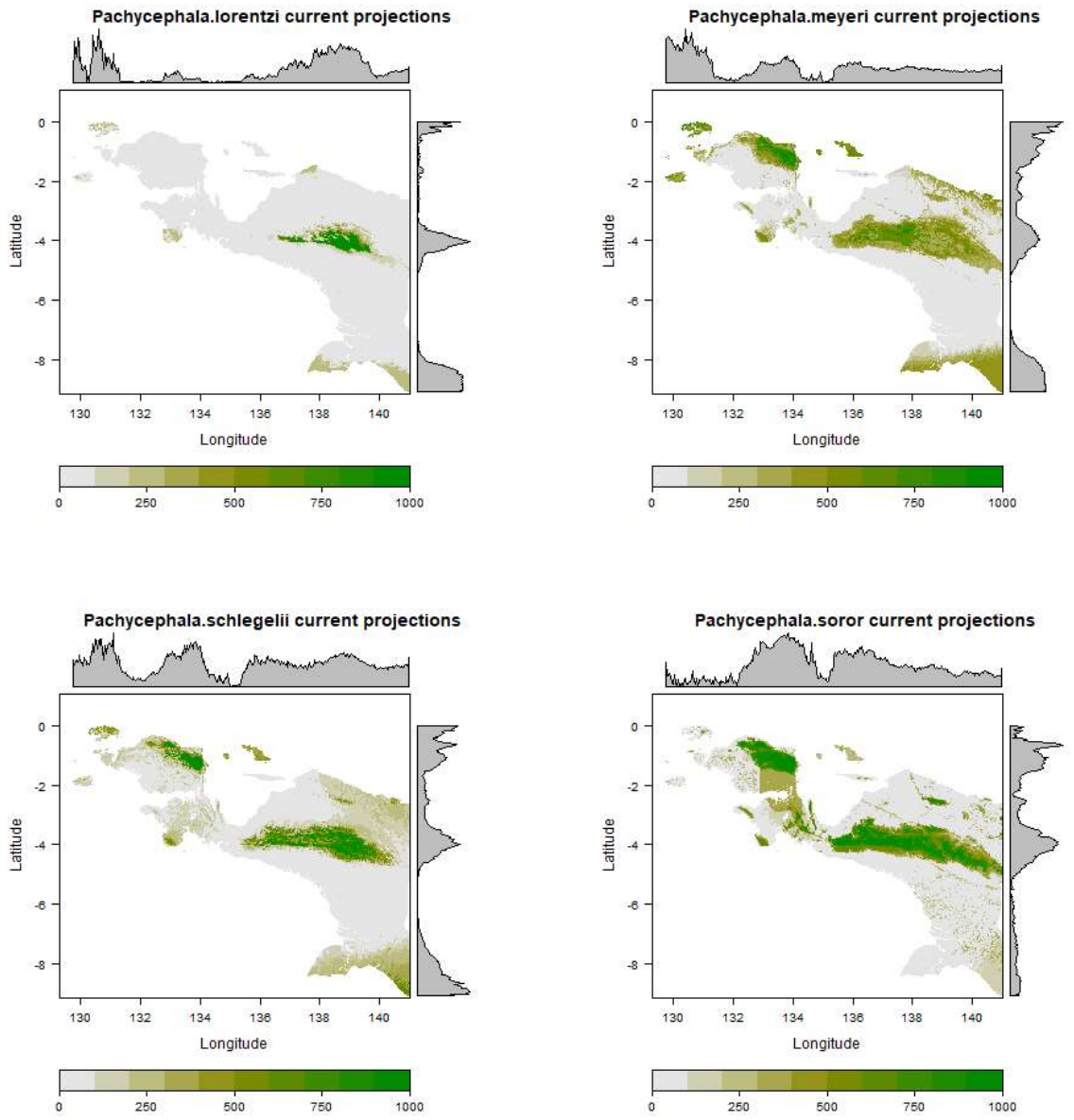
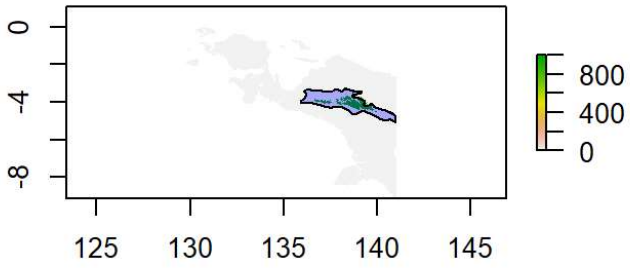


Fig. X. Ensemble plots for the four species; Greener color means higher confidence that the birds occur in that area. The charts to the top and the right show a relative amount of how much of the land cover in that height/length are covered with predictions.

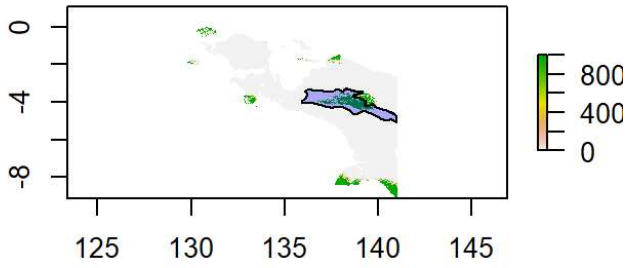
3.5. Comparison SDMs and IUCN

Overlaying the calculated SDMs with the range maps received from IUCN birdlife shows varying degrees of overlap, depending on species and model used. Some example cases include the GLM (and the very similar looking ANN) for P. lorentzi, which overlap completely with the IUCN range data, only the modelled area is a lot smaller than what was expected in the birdlife data. In other cases, such as P. meyeri, all of the models predict a much larger range, specifically also in the central mountains, than IUCN which shows they only appear at the northernmost area of Papua Barat. P. schlegelii and P. soror Birdlife populations are fragmented, which is als the case for the modelled SDMs, only it is very hit and miss where the areas overlap and where they don't.

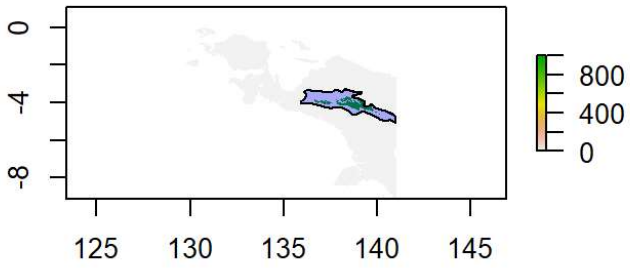
Pachycephala lorentzi GLM



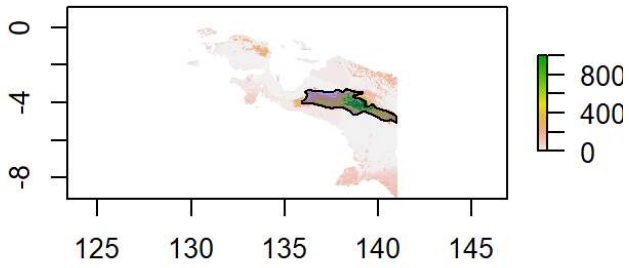
Pachycephala lorentzi GAM



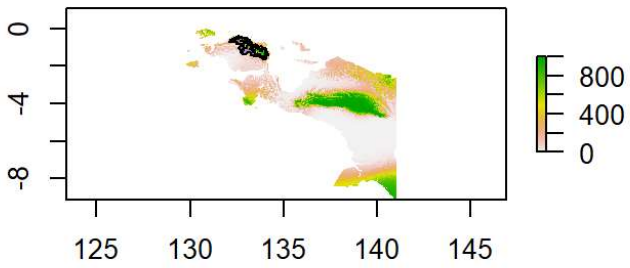
Pachycephala lorentzi ANN



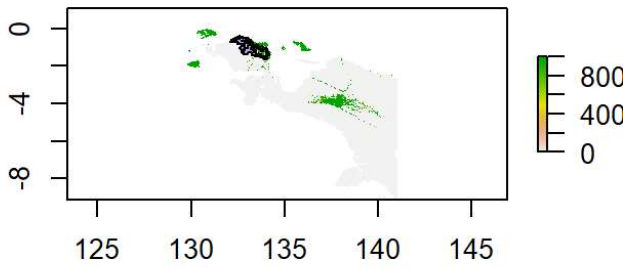
Pachycephala lorentzi RF



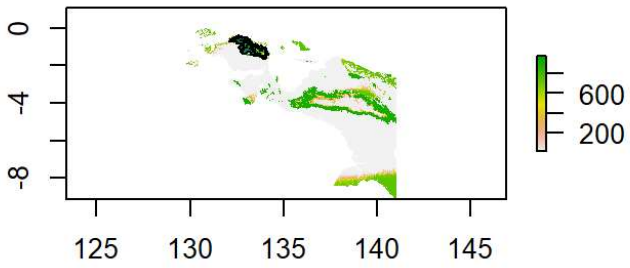
Pachycephala meyeri GLM



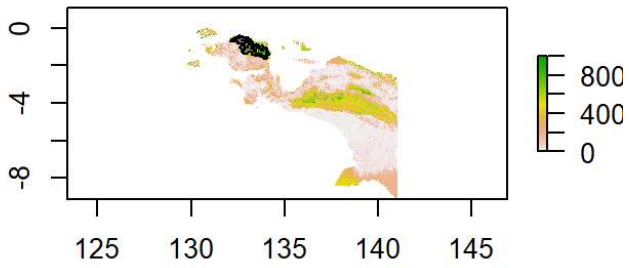
Pachycephala meyeri GAM



Pachycephala meyeri ANN



Pachycephala meyeri RF



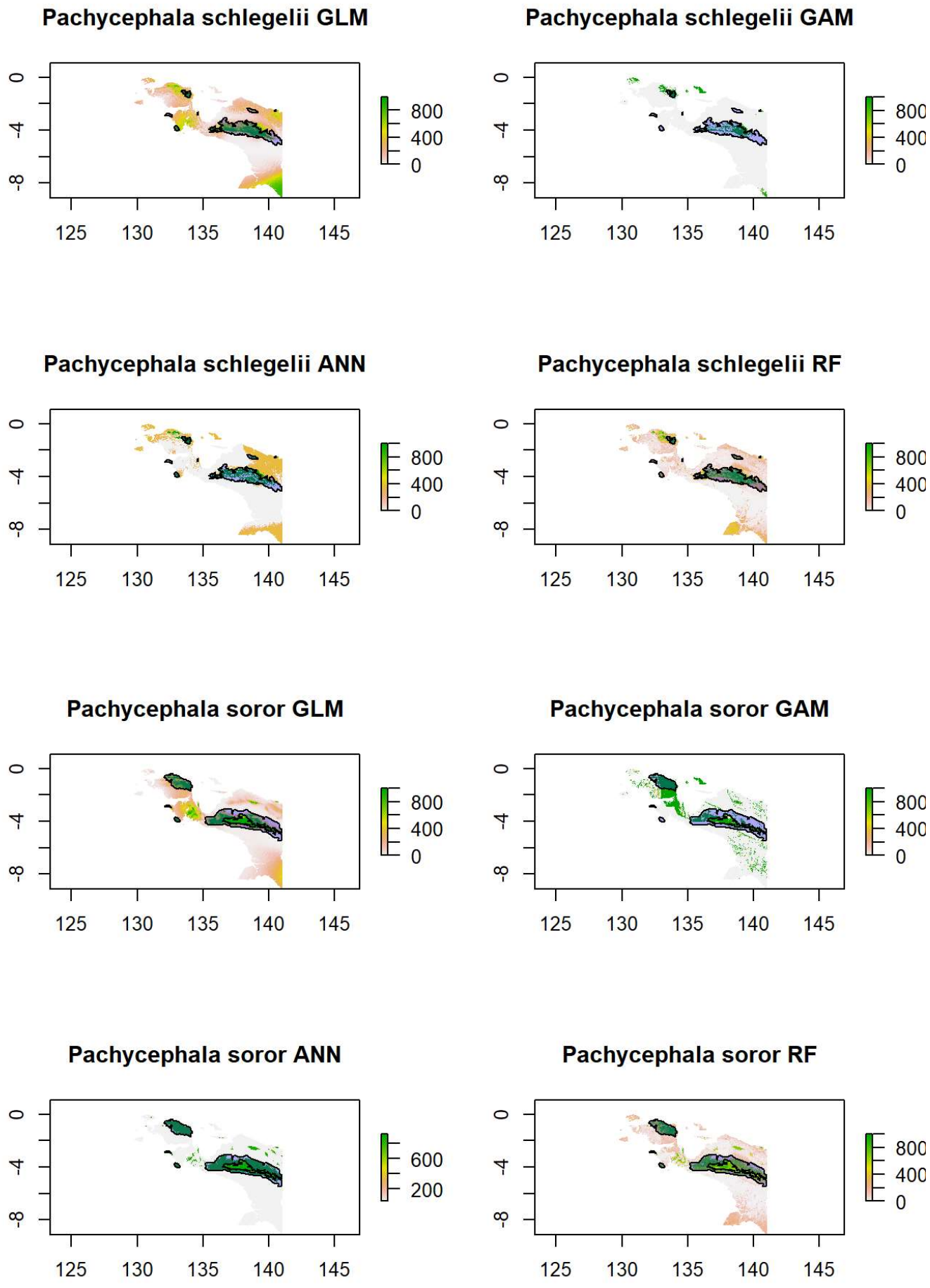


Fig. X. The calculated SDMs GLM, GAM, ANN and RF are overlaid with the light blue IUCN range maps.

3.6. Comparison ensemble SDMs and IUCN

The Ensemble/IUCN comparison is analogous to the previous section. The depictions are a combination of the other SDMs.

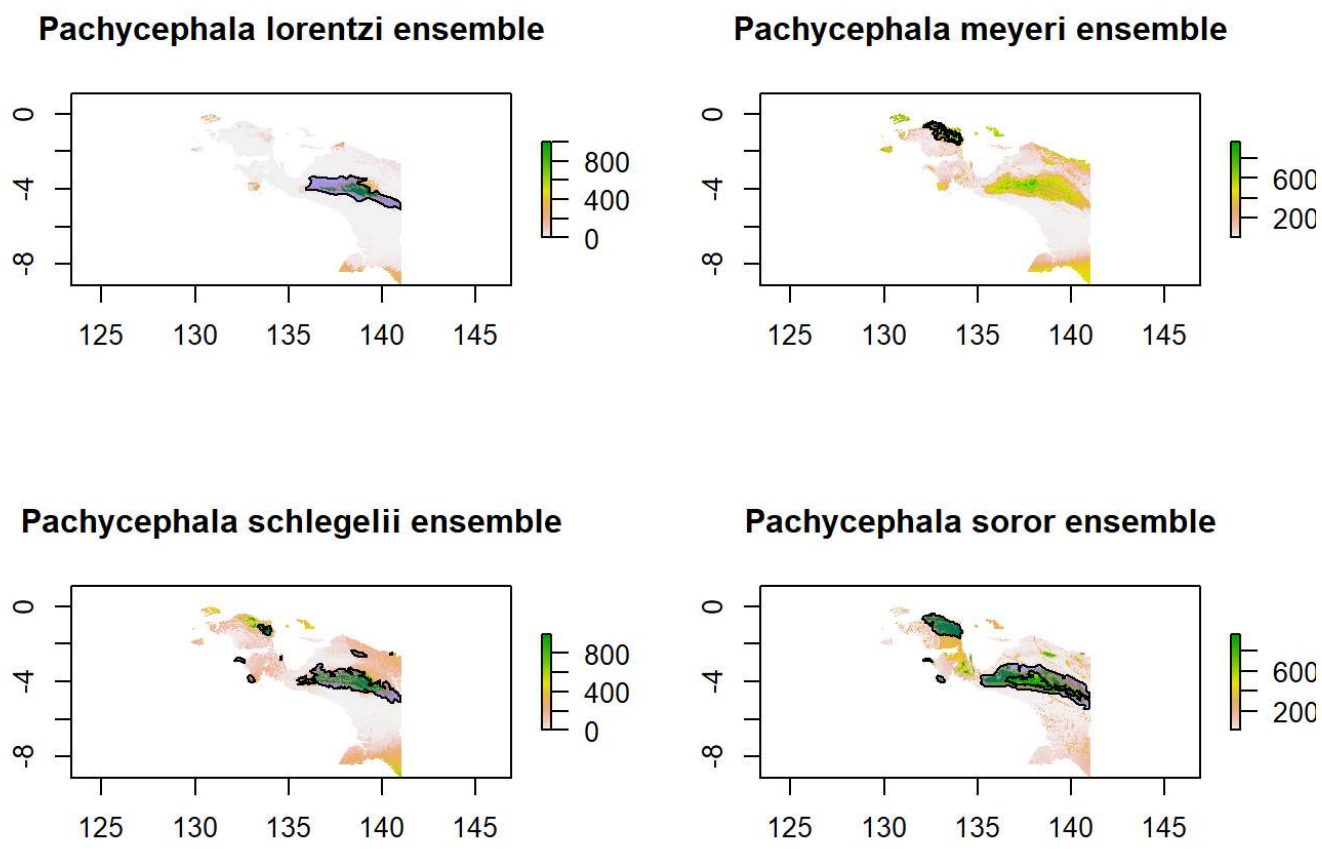


Fig. X. The ensemble models are overlaid with the light blue IUCN range maps.

4. Discussion

Data validation How many percent of the present data, How much of the habitat is covered by the SDMs? Why is the SDM realsitic/unrealistic? → where do the food resources occur etc. Validate the data sources (less points, only in areas with high population density….) Warum manche keine einfluss?

What can be seen in the results?

Looking at the results, while there is noticeable overlap and good tendency, especially in the mountain ranges, the modelled ranges generally differ from the IUCN birdlife ranges in multiple areas. Reasons for this discrepancy lie in some possibly important indicators missing, like food sources or geographical barriers. Another likely reason is that some fine-tuning of the models could be valuable to deliver more accurate results. Sehr wichtig: Bias of the areas were data was taken!! → nur einzelne punkte, zb sehr relevant für höhenmeter. “We use a common Eurasian butterfly (Aglais urticae) as an exemplar taxon to provide evidence that range model quality is decreasing due to the spatial clustering of distributional records in GBIF. Furthermore, we show that such loss of model quality would go unnoticed with standard methods of model quality evaluation.“ (Beck et al., 2014). https://www.sciencedirect.com/science/article/abs/pii/S1574954113001155 (https://www.sciencedirect.com/science/article/abs/pii/S1574954113001155) Also include: fressfeinde, nahrungskonkurrenten, food resouces, symbiotische arten, Zb wo kommt der baum vor As there are many

available options, i.e. in biomod2's own BIOMOD_ModelingOptions(), the use of validation data, more sophisticated creation of pseudo-absence data or even inclusion of real absence data, surely some improvements can be made.

On ensemble models: https://onlinelibrary.wiley.com/doi/full/10.1111/ddi.12892 (https://onlinelibrary.wiley.com/doi/full/10.1111/ddi.12892) "A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD" "The idea of combining predictions from different models into an ensemble has gained considerable popularity in species distribution modelling, partly due to free and comprehensive software such as the R package BIOMOD. However, despite proliferation of ensemble models, we lack oversight of how and where they are used for modelling distributions, and how well they perform. " (Hao et al., 2019) One could argue that in the case of a suboptimal model, the results get worsened by being put together.

5. Conclusion

The biomod2 library in R is a powerful tool to create with easily modifiable settings a range of different SDMs, given just some species presence data and a range of indicator raster files. Using said library multiple SDMs were created for four bird species in Indonesia based on GBIF occurrence data. These SDMs were compared to IUCN birdlife data. Some overlap in the modelled and test distribution ranges was found, however a lot of at times unrealistic results were created. Different elements like bias in the selected data, lack of possibly important predictors and options to change are important places to consider improving. There are many ways imaginable to expand on this work.

The SDMs have some overlap, however more research is needed to get towards more trustworthy SDMs.

References

Griffith, D.M. (2017) Species Distribution Modeling in R. 15.

Appendix

Images

… your images here

Code

MCMMB_Main.R

Code

species_loops.R

Code

presence_absence.R

Code

Indicator.R

Code

sdm_biomod2.R

Code

bird_plot_loops.R

Code