Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2016-65-AC1, 2017

Earth System Science Data Discussions

Interactive comment on "A Global Model of Predicted Peregrine Falcon (*Falco peregrinus*) Distribution with Open Source GIS Code and 104 Open Access Layers for use by the global public" *by* Sumithra Sriram and Falk Huettmann

Sumithra Sriram and Falk Huettmann

sumithrasriram@gatech.edu

Received and published: 8 June 2017

1. Generality of the model and the transferability

Our reply: Thanks, this comment came from the public and it needs to be said that our model is already global, thus it generalizes worldwide (more is not possible, than a global inference)! The model presented here is a generic one that generalizes across a compilation of diverse predictor layers. All available global peregrine falcon data were used as an example to illustrate the use of the unique data cube compiled, and the case

study assessments clearly show for everybody to see that the model obtained is not only accurate, but also highlights new, important and so-called unconventional factors not previously known influencing the distribution of these species in the Anthropocene. This is new scientific information and a template on a global scale, to be extended to any other terrestrial species for obtaining a general and unbiased distribution model for better conservation management. We see no relevant bias or lack of generalization in that. Some of such work was already started by us earlier, too (e.g. Huettmann et al. 2011, Kandel et al. 2015). Thus, our model is already fully generalizable, and applied, worldwide, to any pixel in the world! We feel, the author of this comment did not really understand that concept of global model predictions and such an inference (as we argued above).We showed in the maps assessments about how well those models perform and generalize (98 % AUC), globally. This is virtually unachieved other than this study. So we think we have addressed this comment in our work to good satisfaction.

2. The source data are non-representative

Our reply: Thanks, but that cannot be so correct because we use best-available data world-wide; that is true for the 100 (!) GIS habitat predictors (all describe the habitat pixels where falcons occur), and for the 60,000 falcon observations worldwide (these are sites where the falcons where actually seen, presence; unbiased). These are the best data at hand for such work and this species ever used! So we have used all the available data that were available and see no bias. We discuss all relevant details of these topics in our manuscript. We think this comment tries perhaps to speak to the fact of survey effort; and normal distribution, parametric assumptions. We have clearly shown and stated in the MS text that these are smaller issues in our work, for instance due to the high AUC assessments and non-parametric data mining algorithms that can account for parametric assumptions. To further help this reviewer and comment, we suggest to read work done by Kadmon et al. (2004) regarding so-called road biases and why this is no problem in the methods and approaches we use. The use of proxy

layers is widely used, and indeed, we use such predictors to great success. This is well known and applied for decades now. So we see no issue on non-representativeness here and we feel that our text speaks to those facts fine.

3. The authors created a situation in which country-specific socioeconomic factors can erroneously be given explanatory power, whereas environmental variables that biologists know to be important in affecting peregrine distribution are not included.

Our reply: Thanks, but peregrine falcons are migratory and cross many nations. This is not caused or created by us. We used 100 globally consistent GIS layers designed for global analysis, but we used the actual pixels for our analysis scale. So we do not have a national bias or artefact; nobody has ever used a global pixel-based analysisfor peregrines. So to argue we have ignored predictors, or we would run a national bias, sounds very odd to us, and when a global analysis is the aim. It's just not true. Our GIS layers act as direct, and proxy predictors along the flyways and ranges of this species. We wish to add more, if we are told which ones that would be (the reviewer left it as an unsubstantiated statement). The models simulated were un-biased, impartial and run without any intentions other than to obtain best possible predictions for inference (that's Leo Breiman 2001, basis of machine learning). All 104 layers, including commonly used environmental variables, were included in the analysis. Repeated runs of the models consistently picked the variables mentioned in Table 1 to be the most influential factors that determine the predicted distribution of these falcons. We speculate that this comment might be driven by the idea that falcons would just occur in the 'wild', and are not urbanized. Our models, and all recent literature shows entirely the opposite for a global analysis and for this species. This is easy to show, as provided in the MS text, e.g. release of hand-raised habituated individuals while the wild species DNA/stock became extinct and urbanization on the rise. We still have (sub) samples from virtually all areas of the globe in our assessment; our point sample map shows that clearly!

4. The resulting model does not appear to fit the testing data well – cutoff point of 0.01

Our reply: Thanks, we think the reviewer does not really understand that we use 'classification trees', recursive partitioning, but not logistic function 0 to 1 which is symmetrical. Trees are none of that! It's the fundamental difference from frequency statistics; our methods are widely published, e.g. see Kamel et al. 2015 and Mi et al. 2017 etc. The cut-off is non-symmetric, rightly so, and simply stems from the assessment with alternative testing data. This is done in machine learning for the best predictions, and the best inference. Whereas the 'model fit' (like an r2) is 100% not relevant in that, as per Leo Breiman 2001 (inference from predictions, but not from model fit requirements). The ROC curve of the model chosen shows an accuracy of 98 (2% error, globally!)%. That means the data are to 98% reproducible, worldwide; it's a world record and certainly for this species! The RIOs of all the present points were extracted, and to accommodate the 98% accuracy error rate when compared with real data! Thus, 2% of the points on the lower side of the RIOs were considered to be erroneous (easily >95% of the alternative data correctly classified; beyond significant!) and this results into a threshold of 0.01. Such high accuracy hints at an almost perfect recreation of the training data set, and to be sure and spatially precise, other validation methods were performed too for a better proof of evidence. The testing data was obtained from an independent source, and when the RIOs w.r.t the model were extracted, it was observed that less than 2% of the points had a RIO of less than the threshold value, which again, adheres to the 2% error assumption. It's an alternative test matching from what we initially stated. These methods are widely done and published, e.g. Elith et al. 2006, Huettmann et al. 2007, Kandel et al. 2015.

5. The predicted map in Figure 3 does not appear to correspond well with some other smaller-scale peregrine distribution maps.

Our reply: These matches exactly because they are based on each other. Figure 3 is the base raw heat-map obtained from the model and shows the relative index of occurrence (RIO; no threshold). The actual distribution map interpreted from this model for peregrine falcons is shown in Figure 7 and it is EXACTLY based on Figure

3 and its structure in the legend and pixels. So we kindly disagree with this comment by the reviewer. We are happy to send over both maps for the check; as needed. Here we offer in our paper the raw model as well as the interpretation, and then we show how well those work by using alternative data (as outline in previous reply points and citations we provided for the evidence).

6. Combining different types of data -breeding, non-breeding, and migratory locations

Our reply: Right, this is exactly what was done. The paper mentions that the model provided for these falcons is general and global, and includes all the areas that are suitable for these falcons year-round. We refer to the GLOBAL ECOLOGICAL (YEAR-ROUND) NICHE. That's exactly our aim and achieved here with a very high accuracy. We fully agree that a nesting niche, or a wintering niche is located within the text should be very clear on this.

7. Give units for pixel size column in Appendix A.

Our reply: Thanks, we changed it in the manuscript. Thank you for pointing out the error.

Changes in the manuscript: Updated Appendix 1, "Original pixel size" column, with units.

8. How can the same model be fit to mean October and November temperatures in both the Northern and Southern hemispheres?

Our reply: Thanks, we run a global model, so we use global predictors for every 1km pixel on earth. Sure there are different processes and things going on by hemisphere and season etc. However, our 104 predictors catch that fine, certainly the global climate model predictors. We do infer from the predictions, explicit in space and time. Thus, what the reviewer describes is 100% not a problem (this would be a potential problem when inference is done from parametric assumptions and model fits; but that's exactly NOT what we do). Our predictors describe the pixels and their climate explicit in space

and time (see Worldclime for an example, or any other global model; let's say Wei et al. 2011)

9. Explain specifically which data were selected and how they were screened.

Our reply: The raw collection of data points included almost 500,000 points, of which duplicate records, records with incorrect geo-referencing and records with ambiguous data were removed. Also, only points that were reported after the year 1990 were taken into consideration. Our process results into a conservative set of valid peregrine sightings. We are happy to send the reviewer the raw data and our test set. All of this is fully transparently done based on the GBIF data set.

Changes to the manuscript: (Section 2.1) "This raw data had to be filtered for accurate and duplicate records, for records with incorrect geo-referencing and for records with ambiguous data to finally obtain 60,261 unique presence points, less than half the size of the initial raw database."

10. Check Figure 10, a visual comparison Figure 7 to a map WDPA protected areas suggests that many protected areas do appear to be in areas with predicted occupancy by peregrines.

Our reply: Thanks, we can let others tosee and decide well. Please have a look: It's pretty clear that the protected areas are not where the prediction hotpots are, and our geo-referenced overlays (done in a GIS thus reliable) show just that, without relevant margins of errors. The trends are pretty clear. Due to the "synurbanization" of the falcons (which is widely published, citations provided, and explanations given earlier above), the falcons are flocking more towards urbanized areas. This has been captured correctly by the model, and by the reported points in GBIF. Conservation measures for protecting these birds need to be changed accordingly to accommodate this change. A classic example is found in Moscow/Russia, for instance, where the Stalin architecture buildings became a cliff, host for urban peregrine nests, and those birds there are active at nights and feeding on pigeons! Our model shows exactly such things!

11. Although the authors say they use a 1 km x 1 km grid for all of their variables, the text should make clear that many variables are country-specific variables.

Our reply: Thanks, sure, this has been added to the manuscript to be clear. Certain layers like the socio-economic layers are available only at a country and county level resolution for privacy concerns of social scientists. Such layers have simply been sub-sampled for alignment with the rest of the layers in the model and its pixels. So we find we have addressed all of this is fine in the MS text, as is.

Changes in the manuscript: (Section 2.1) “Certain layers such as the socio-economic layers were available only at a very coarse resolution, due to privacy concerns. Such layers have simply been sub-sampled for alignment with the rest of the layers in the model and all its pixels.”

12. Antarctica was removed from the analysis; similarly, the Greenland ice sheet is not potential habitat and also should be removed from the analysis

Our reply: Thanks, our model correctly predicts that the Greenland ice sheet is not potential habitat or part of the ecological niche (while parts of coastal Greenland are; that is well reflected in the species literature). We find, the continent of Antarctica, ice-covered, can be excluded (but we kept the southern Antarctic islands and regions) . So we think this critique is addressed in our work.

Lastly, we think that another key aspect of this manuscript, the delivery of 104 (!) GIS layers for the public all done Open Access free of charge has not received sufficient appreciation by those comments. So we wish to emphasize that further, and beyond ‘just’ the falcon model (which is great by itself and with such a great accuracy, all updated as stated!)

Interactive comment on Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2016-65, 2017.