Part 2 Questions & Answers Session

Please type your questions in the Question Box. We will try our best to get to all your questions. If we don't, feel free to email Juan Torres-Perez (juan.l.torresperez@nasa.gov), Morgan Gilmour (morgan.e.gilmour@nasa.gov) or Claire Teitelbaum (cteitelbaum@uga.edu).

Question 1: I recently conducted my dissertation on Emperor Penguins in Antarctica. I used Sentinel and Landsat imagery simply because of their open access, but I was wondering if you know of anyways to access higher resolution imagery without such massive financial strain
Answer 1: Web platforms (e.g., NASA EarthData) can be a great place to start when searching for remote sensing layers. Most will allow you to filter by spatial extent and resolution. Depending on your funding source, there may also be other options for acquiring commercial high resolution satellite imagery; for example the Commercial Satellite Data Acquisition program can provide access to some data from commercial vendors for U.S.-Government-funded research: https://www.earthdata.nasa.gov/about/csda. Other programs might be available for other countries and funding sources.

Question 2: Can you explain how you captured the frigatebirds, what type of sensors you used, and how many were installed?
Answer 2: We captured frigatebirds by hand at their nest sites. We used e-obs solar GPS tags, and there were 7 birds in this dataset.

Question 3: What do you think about the "amt" package of R? Are better packages or using "amt" better? Is the R script used available to download?
Answer 3: "amt" is a very useful package because it provides a streamlined and customizable way to do many common tasks in the analysis of animal movement data. Many, if not most, of the underlying functions come from other packages, so it could be useful to check out those other packages for more information or more in-depth analysis. There are also a number of other packages that will be useful, depending on the types of analysis you want to conduct! The R script is available in the training GitHub repository: https://github.com/NASAARSET/ARSET_Animal_Tracking

Question 4: Do you have any recommendations for R packages that would be useful for presence-only data with no pseudo-absences? Using Maximum Entropy or other?
Answer 4: The maxent function from the dismo package is useful for presence-only data. Check out the manual page for maxent here: https://search.r-project.org/CRAN/refmans/dismo/html/maxent.html

Question 5: How can I include the altitude/elevation data to my analysis? For example, in a floodplain some areas aren't available for a terrestrial species due the high water level, and I want to analyze it.
Answer 5: Altitude or distance to water could be an additional predictor in your models along with the environmental covariates. You can gather elevation from digital elevation model (DEM) products and water from remotely sensed surface water products.

Question 6: What coding platform do you think is best for this sort of data analysis? Is R Studio the only option? What are the strengths or limitations of the R package amt for analyzing animal movement compared to other packages? / Is there a particular reason to use R instead of Python? Could the exercise be repeated with Python?
Answer 6: We are both primarily R users, which is why we used R for this demonstration. Theses analyses could almost certainly be performed with Python. R is the most commonly used language in the ecology community, so you will probably find more packages and tools for R than in Python, but Python is also set up well for geospatial analysis. See Question 3 for the answer to your question about “amt”!

Question 7: Is there any problem having NAs in my data? If there is, what should I do?
Answer 7: There are many ways you can deal with NAs in your data. If there aren’t many, you will usually just omit those points from analysis. If you have too many NAs, you might consider using another data source (for example, a lower-resolution raster layer) that has less missing data where your animal is tracked. If you know why the data are missing (for example, water depth is missing for a seal when it is on the beach), you could impute values.

Question 8: Is there a reason why sp and raster are being used instead of sf and terra in the for loop to extract cl a or bathy?
Answer 8: The R-packages sp and raster are one example of how we can append the environmental variables to the tracking data; sf and terra are definitely other options!

Question 9: In the work flow of importing and extracting remote sensed datasets, projections weren't mentioned... At what point in the process do you recommend re-projecting datasets? Could you do that with the .netCDF file before extracting?
Answer 9: It's great that you're thinking about projections! Environmental data are typically available/downloaded with latitude/longitude coordinates, but you can change the coordinate reference system for your analyses and for plotting within R. CRS is a function from the R-package "raster." Lambert equal area projection can be helpful depending on the size of your dataset.

Question 10: How are boosted regression trees different from random forest?
Answer 10: Boosted regression trees use a linear combination of multiple trees to come up with a final model. This method allows poorly fitting trees/models to receive lower weights in the final model, which usually improves model fit and performance. In BRTs, trees are trained sequentially, so each tree can attempt to correct errors in previous trees. In RF, trees are fit independently, each on a subset of the training data. In general, RF is faster to fit, and BRT is more accurate - though this can depend on your data and model.

Question 11: If you generate pseudoabsences from random locations in a SDM, what makes it conceptually different from a SSF?
Answer 11: You could consider an SSF a type of SDM, with a more constrained definition of pseudoabsences.

Question 12: What type of model (GLMM, GAMM or Boosted regression tree)is better in each context?
Answer 12: The type of model really depends on your tracking dataset and the types of relationships (e.g., linear; non-linear) that you might expect with the environmental covariates. Plotting a subsample of your tracking data with the environmental variables that you're interested in can give you a sense of the shape of these relationships. You can also try running the different model types on subsets of your data to assess which models perform the best.

Question 13: Also, if I have erroneous points due to punctual errors in the moment when my data was taken? Can I work with that or I need to erase them? If I need to delete them, is there a way to do it with R or do I need to go one by one?
Answer 13: It's ok to omit points from your dataset, and applying a speed threshold, for example, is one way to do this. You can also set geographical limits on your locations and exclude all points outside that box.

Question 14: Im new to R so im scared of the warnings. I saw that in some lines it gives you warnings that say "removed 300 lines." What does it mean? Is there a problem with that kind of messages because in my data I get some of them too.
Answer 14: Warnings from R may simply refer to how R processed your code or displayed your results. It may also mean that there were 300 NA's in the data, for example. It doesn't necessarily mean anything is wrong, but it's important to regularly inspect the dataframe and code you're working with to make sure it looks like how you expect.

Question 15: What is the conceptual difference between step selection function and resource selection function? They are usually presented next to each other in the literature, but I am not sure I understand the difference between the two approaches...
Answer 15: A step selection function is a type of resource selection function. A resource selection function is any statistical model that estimates the relative probability of use of a resource unit compared to alternative (non-used) resource units. An SSF compares used resource units to available resource units based on movement characteristics of an individual animal.

Question 16: As MODIS data are at quite large spatial scales, are they appropriate to use for extracting environmental covariates when considering movements of species that don't have large home ranges?
Answer 16: If an animal is only moving within one or a few pixels/grid cells, then you are unlikely to learn much from remote sensing data at that scale.

Question 17: Why did you choose MODIS NDVI that is 250m resolution rather than Landsat NDVI data that is 30m resolution?
Answer 17: Landsat data would be another good (or better) choice in this example! In this case we used MODIS because it was easy to download from NASA EarthData and

show as an example, and because it required less processing. Because MODIS data are collected more frequently at lower spatial resolution, the composite images can have fewer issues with cloud cover, which is another important consideration in addition to spatial and temporal resolution.

Question 18: In the construction of SDM, are normalized covariate variables required?
Answer 18: Sometimes it's helpful to center and/or scale the environmental covariates, but it's not necessarily required.

Question 19: Is it possible to use these R codes to develop models in climate change scenarios to project species distributions?
Answer 19: Yes, methods like these can be used to project species distributions under climate change scenarios. If a model is fit using current climate data, it can be used to make predictive maps. One example of this approach was conducted with data from Palmyra (demonstrated in the SDM section) in this recent study: https://doi.org/10.1111/gcb.70138.

Question 20: Why is the same step_id for several locations? wouldn't be just 2 points first and last?
Answer 20: The same step_id was present for a used step and its paired random steps. Each used step has a unique step_id, but many random/available steps can have the same step_id.

Question 21: Regarding Species Distribution Model: Is there literature to help us decide on which method to use to create pseudo-absence and which type of predicting model to use based on our data and research question? Or do you recommend trying every method and comparing the fit of all the resulting models?
Answer 21: This is an area where there is a lot of literature comparing methods. It is probably not possible to try every method, because there are so many and each one can be computationally intensive. For example, the paper mentioned in the SDM section describes and compares a few different methods with tracking data from an elephant and a blue whale: https://doi.org/10.1186/s40462-021-00240-2. Another example: https://doi.org/10.1111/j.2041-210X.2011.00172.x.

Question 22: Hi, within step selection functions, can you impose limits or conditions on the simulation of control steps that are study system specific ? e.g., simulating steps for a seabird that would not fly over land (even if the end point is at-sea)? Or might that suggest step selection is a less suitable form of modelling in this instance? Hope this makes sense, thank you (and thank you for all of the training!)

Answer 22: Yes, this would be possible. Using the code in the demonstration, one way to do this would be to simulate more control steps than you need, extract environmental variables along each step/path, and exclude any control steps where the segment crosses land (or ends on land). In this specific example, it's also important to note that SSFs do not model the actual movement trajectory between the start and end points, so you would not necessarily know whether a simulated step would require the seabird to fly over land, just whether a straight line connecting the start and end points crosses land.

Question 23: Does rastering help get clear images every time? Even in the bad weather?

Answer 23: It's sometimes helpful to take the mean value across images in order to obtain the best environmental data. This is why using coarser time resolutions, like weekly or monthly data instead of daily data, can be helpful.

Question 24: Could you elaborate again why choose to obtain env. variables at the end step point, rather than start step point?

Answer 24: Paired used and available steps share a starting point (see the schematic in the slide deck). They will therefore have the same environmental conditions at their starting points. Conditional logistic regression models step type (used or available) as a function of the environmental variables, so using only environmental variables at the starting points of the step would mean that all pairs have the same x-variables and the model would have no variation to work with. However, you might consider including conditions at the starting point as an *interaction* with conditions at the end points of the steps. For example, this could allow you to ask if an animal is more likely to move to a certain habitat type when they are starting at a warmer vs. a cooler location.

Question 25: How do you deal with the "burst" column in the SSF model? Or did you use it in the demo?

Answer 25: The burst column is used internally in the steps_by_burst() function in “amt.” When constructing steps from a track, “amt” ignores steps that “cross” bursts (e.g., start in burst 1 and end in burst 2), because these steps would have too long of a time interval. This ensures that the distributions of step lengths and turning angles are based on regularly sampled data (when using the random_steps() function).