Questions & Answers Session Part 1

Please type your questions in the Question Box. We will try our best to answer all your questions. If we don't, feel free to email Amber McCullum (amberjean.mccullum@nasa.gov), Juan Torres-Pérez (juan.l.torres-perez@nasa.gov) or Zachary Bengtsson (bengtsson@baeri.org).

Optional For Part 2: Although NOT A PREREQUISITE, in Part 2 there will be a demonstration of Species Distribution Models in **Wallace,** an ecological modeling application. If you wish to follow along with this demo, before Part 2 follow these instructions for download and installation.

Wallace v1.1.0 Instructions for Installation:
http://appliedsciences.nasa.gov/sites/default/files/2021-07/Wallace_Installation_V1.pdf

We will have the recording of this demonstration available within 48 hours after the presentation for you to go through the demonstration at your own pace.

Question 1: How unique should the species' environmental envelope be defined? In the case of tropical forests where distribution is more or less the same for several species, how do you suggest species be differentiated?
Answer 1: The environmental envelope for multiple species, especially in the tropics, will be similar. The environmental or climate envelope delineates the areas that have suitable habitat for a particular species. Because the climate layers and species ranges can be similar, the modeled suitable habitat might be very similar. I would suggest use of extensive occurrence data for the particular species of interest. While the climate/predictor inputs may be similar for multiple species, the occurrence datapoints will assist in differentiating the distribution of different species. I would also suggest modeling different species independently. Here is a nice resource for climate envelope modeling:
https://crocdoc.ifas.ufl.edu/projects/climateenvelopemodeling/publications/Use%20and%20Interpretation%20of%20Climate%20Envelope%20Models%20-%20A%20Practical%20Guide.pdf

Question 2: Like Maxent, which is also available inside Wallace, it is a presence-only SDM. How accurate it is compared to models which consider both presence and absence? Is there any presense-absence SDM model which have Graphical User Interface (GUI) like Wallace?
Answer 2: Maxent is one of the most extensively used SDM models, and has a generally high level of accuracy for many species. With absence data, there are a couple important considerations I have outlined on slide 37. SDM such as Maxent or GARP, referred to as presence-only methods, actually do require the use of background data or pseudo-absence data. As confirmed absences are very difficult to obtain, especially for mobile species, and require higher levels of sampling effort to ensure their reliability compared with presence data. Presence-absense SDM have been shown to be highly accurate. Here is a great resource for a comparison of presence/absense and psudoabsense (also on slide 37):
https://besjournals.onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2011.00172.x

Question 3: What are the best practices for handling environmental variables that have different resolutions?
Answer 3: There are a couple of options here. You can rescale the data to align with the coarsest resolution data, this is probably the simplest approach. You could also conduct downscaling on the coarser resolution data. There are algorithms you can run for each of these approaches.

In general, you want to indeed re-grid and conduct the comparison with both data sets at the same resolution, and preferably, the re-grid would use a similar or same gridding scheme. The coarser grid set is normally the limiting factor, unless you have the original data so you could regrid the data to a finer resolution.

Here is a resource for upscaling:
https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.13301
There is also a rescaling function in ArcGIS, which is not something I have used but looks promising: https://desktop.arcgis.com/en/arcmap/10.3/tools/data-management-toolbox/rescale.htm

Question 4: Is MRLC, or some variation of MRLC, available for other countries?

Answer 4: MRLC is only available for the US. However, there are other resources that provide land cover products on a global scale. One example of this is the CCI land cover product displayed here from the European Space Agency: https://maps.elie.ucl.ac.be/CCI/viewer/

NASA's MODIS land cover product is a coarser dataset, but is globally available. Land cover can vary by region and global land cover products are good for classifying different land cover types, but will not account completely for variations and uncertainties within your area of study. Check out our training using Google Earth Engine for Land Cover Mapping.

Question 5: Is it possible to do species distribution modeling with few occurrence data? Like with species location data less than 10?
Answer 5: The more you have, the better, but many species distribution models can work successfully with as **few as 10-15 presence points**. I would also suggest looking into finding more presence data points in some of the data portals we mentioned like GBIF. Conducting multiple modelling efforts across multiple models can also be effective.

Question 6: How authentic are the species distribution records are on GBIF? How these data are recorded?
Answer 6: There is a strict set of data standards for GBIF. Participants and publishers applying shared rules and conventions to describe, recording and structure thousands of different datasets drawn from hundreds of institutions around the world. Common standards are the main enabler for bringing together the hundreds of millions of primary biodiversity records in the GBIF index.The Darwin Core Standard (DwC) offers a stable, straightforward and flexible framework for compiling biodiversity data from varied and variable sources. Depending on how much information the source data contains—and how much they wish to share—publishers can create a Darwin Core Archive with one of three cores:

- a Taxon core, which lists a set of species, typically coming from the same region or sharing common characteristics
- an Occurrence core, which lists a set of times and locations at which particular species have been recorded

- an Event core, which lists field studies (including the protocols used, the sample size, and the location for each).

More information can be found here: https://www.gbif.org/standards

Question 7: How many environmental variables should I use? How would I know if my environmental variables or data points for each variable are sufficient?
Answer 7: This will largely depend on your species of interest and the complexity of the environment within your study area. Generally, the more data available, the better the model. The strengths of one data type can offset weaknesses in another. Combining and integrating different data types often improves overall species distribution estimates. To examine whether or not your data is sufficient for the model, you'll want to complete accuracy assessments and validations on your model. We'll talk more about this in the next session.

Question 8: How to evaluate the 'sampling adequacy'? Is there any method to quantify the uncertainty of presence/absence data?
Answer 8: You can evaluate the fit of models using different occurrence points. There can often be bias in the sampling in regions where it is easier to obtain field data for example. Its also important to remember that all models are "wrong" but useful! You can evaluate the models using things like area under the receiver operating characteristic curve (AUC)and correlation (COR). AUC evaluates how well model predictions discriminate between locations where observations are present and absent,
and is one of the most widely used threshold-independent evaluators of model discriminatory power.You could also look at the standard deviation within distributions to assess the presence/absence data.
Here is a great paper on sample size and model performance:
https://www.researchgate.net/publication/235957626_Effects_of_sample_size_on_the_performance_of_species_distribution_models

Question 9: How do you reconcile the mismatch between satellite measurements e.g. Temperature and Insitu measurements. I see most SDMs relying on satellite data but haven't seen someone accounting for these sources of error.
Answer 9: Satellite products typically go through some sort of quality control step to eliminate pixels with values that clearly do not represent ground conditions, and many products provide disclaimers of potential sources of error. Values obtained from

satellites are also validated with in situ data in many cases. You will need to look into the exact data specifications of your chosen datasets to see what steps may have already been completed to account for these mismatches.

Question 10: In the case of big migratory animals like elephants, they may tend to adopt forest trails for ease of movement, which may lead to over abundance of presence data near trail areas and the model may show false distribution patterns. Can this and other various behavioural tendencies be incorporated while designing the model?
Answer 10: Yes! Good point. Animal movement and migratory patterns are very important when modeling suitable habitat. We will discuss this in further detail in Part 3. One of the tools we will cover is called Circuicscape which helps in mapping migratory movement patterns.

Question 11: Numerous previous studies have shown that empirical statistical models exhibit high accuracies if there's sufficient reference data available from the study area. The models are/could be site, species, or even time specific. This limits transferability. How do you guys or your model(s) address the issue associated with empirical model transferability?
Answer 11: We typically focus on species distribution modeling platforms that act as a general framework to base our work off of. For example, the Wallace platform provides a step by step workflow for creating an SDM, allowing the user to alter environmental and occurrence data depending on their focal species. Input data will need to change from species to species, since presence and habitat suitability varies. However, the framework and procedure for creating an SDM can be standardized within a chosen platform. Make sure to attend our next session for an exercise using Wallace!

Question 12: When would you recommend running multiple models for the same dataset? Also, what would be a good framework to select one method over the other?
Answer 12: This depends on your modeling method of choice. For example, if you choose a machine learning based approach, the model will be run iteratively many times to produce the best results. It's not uncommon to run any model many times to obtain the best result, but you will want to validate and assess the accuracy of your model to determine its performance.

Question 13: One of the major problems I am having is mapping hunting pressure as a predictor variable. The species I am studying are highly hunted, and how to map such pressure is not easy. There is no good spatial data for such human behavior. Do you have any suggestions to overcome such a problem?
Answer 13: This is a little out of the scope of this training, but you might want to look at proximity to human settlements or know recreation areas where hunting is common. These could act as proxies for hunting pressure. You might also want to check with local management organizations to see if they have additional data that might be of use to you.
In this research, hunting pressures were quantified as a function of distance to hunters' access points, human population density, and body size of the species, which are major determinants of hunting impacts.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7540261/

Question 14: Do we have any application on the Marine coastal ecosystem such as presence/absence of seagrass. Are there any pool of data for Marine Coastal domain which are similar to what you presented for the terrestrial domain?
Answer 14: A number of SDMs have been used for marine environments. For example, ENFA has been used to model deep water corals. For seagrass meadows, Maxent has been used. Here's a recent paper from Bittner et al (2020) that used Maxent for seagrass communities along the Texas Gulf coast:
https://www.sciencedirect.com/science/article/pii/S027277141930825X
Here's also another approach from Grech and Coles (2010) who applied SDMs to map seagrasses in the Great Barrier Reef in Australia: https://doi.org/10.1002/aqc.1107

Question 15: Can we get information on bioclimatic data for every single year (past, present and future for different time horizon)?
Answer 15: The temporal and spatial availability of really can vary depending on the dataset of interest. You might first want to identify the most likely important inputs to you model based on what you know about the species. I would suggest viewing the climate data available from NOAA and NCAR (see the links on slide 30). For example, PRISM has temperature and precipitation data from 1981 to present. For mapping future distributions under climate change, you can modify the temperature/precipitation mean and standard deviation depending on what the climate model tendencies in your region, then you can re-run your SDM. Take a look at our

past scenario modeling training (link on slide 32) and the downscaled climate models on NEX (link on slide 33) for more information.

Question 16: Is there a precaution to be taken in mapping the distribution of marine species? Is there any difference between SDMs for terrestrial and marine environments other than adding marine-specific environmental variables e.g. salinity, depth
Answer 16: That would depend on the type of marine organism(s) you are looking at. In general, independent of whether the organisms are sessile or if they swim, there are certain physical oceanographic parameters that are typically used, such as temperature, salinity, etc. For moving animals like fishes, other parameters like ocean currents might have more weight in predicting distribution. For organisms which are typically sessile (corals, seagrasses, algae, etc.) water depth, nutrients, and light availability might be more important to consider when applying an SDM for population analyses. The use of SDMs for marine ecosystems has in general been under-utilized but is becoming more common in recent years. Here's a paper from Robinson et al (2011) where they compared typical methods used for terrestrial ecosystems and how these were applied to marine ones: https://doi.org/10.1111/j.1466-8238.2010.00636.x

Question 17: Many species are highly mobile and can bridge long distances and thus not suitable habitats. How can this fact be taken into account in the models, which goes beyond the concept of false absent?
Answer 17: Animal movement is an important concept to take into account. We will discuss this in more depth in session 3 with Circuitscape. Depending on the species of interest, you could map habitat at different times of the year which could delineate differences in potential habitat locations. You could also exclude regions where the species is migrating through but not spending time in foraging/reproducing/etc. This might necessitate more information about the behavioral patterns of the species.

Question 18: Does presence data have to be direct observations? or can it be known signs of presence e.g. tracks or dung etc

Answer 18: Indirect observations, such as tracks or dung can be used as presence points within SDMs, so long as the caveat is identified and uncertainty with these points are well-evaluated. Depending on the level of confidence of the dung presence points model outputs can be highly accurate. Here is an article where dung was used to effectively model suitable habitat: https://link.springer.com/article/10.1007/s10531-016-1251-2

Question 19: Does the final model comes with accuracy output or will accuracy be calculated elsewhere?
Answer 19: Models are traditionally tested by using half the original species records to build the model and half to evaluate it. Independent data is even more useful when evaluating the accuracy of your SDM. This is systematically conducted by the researcher after the maps of suitable habitat are generated. Depending on the software used, this might be an output, or it might necessitate an additional step.

Question 20: Do you need to exclude predictors variables that are autocorrelated in the construction of the model?
Answer 20: You don't necessarily need to exclude layers that are autocorrelated, but you need to think carefully about which to include. The level of impact on your final output map will depend on the level of autocorrelation of two or more layers. Highly correlated input layers tend to decrease the accuracy of the output. You could test how much this influences your results first by assessing the correlation among the layers prior to input into the models. You could also assess the impact on the inclusion of two or more autocorrelated layers within your model by assessing the accuracy of your output maps with both layers included and then with one or the other layer included.