Questions & Answers Session Part 2

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).

If you have any trouble with Wallace installation or any other issues, please register and send a message to the Google Group [https://groups.google.com/g/wallaceEcoMod], or search the existing messages if there is already an answer to your issue. If your question is about installation, please name your email subject [Installation] for a quicker response.

If you have another question or suggestion, you can also write to the Wallace email: wallaceEcoMod@gmail.com

Question 1: Would any of the resource persons be familiar with predictor variable datasets on global average emergent tree height? I'm asking as emergent trees are a known limiting factor for my species of interest: a tropical forest raptor that depends on emergent trees for nesting. I'm currently using average canopy height as a proxy for emergent tree height.
Answer 1: I'm not aware of any but other variables such as forest structure may be helpful when combined with canopy height for this purpose. More variables may be coming out soon from NASA's GEDI mission or from the GEO BON Ecosystem Structure Working Group, so stay tuned to those projects for updates.

ARSET also provided a recent training on the use of Synthetic Aperture Radar (SAR) data for forest mapping. In session 4, we discussed Forest Stand Height Estimation, here is the training website: https://appliedsciences.nasa.gov/join-mission/training/english/arset-forest-mapping-and-monitoring-sar-data

Question 2: What can the maximum # of occurrences be for Wallace? I have data from 6k-60k and use >10 environmental/biological indicators (through Bio-ORACLE).

Answer 2: If you are uploading your own occurrences, the limiting factor may be the computational power of your computer (i.e. computer memory, RAM). However, to build “good” SDMs you want to ensure you eliminate spatial biases resulting from the sampling design. Wallace allows you to do so with the spThin module within the “Process Occs” component.

Question 3: To what spatial scale is recommended to use R Wallace?
Answer 3: The spatial scale depends more on your research question. Wallace has some incorporated databases such as Worldclim which are available at 30 arcsec, 2.5 arcmin, 5 arcmin or 10 arcmin but using the user-specified module users can upload variables at any resolution (as long as all variables are at the same resolution). Refer to session 1 about having variables at the same resolution.

Question 4: Wallace v2 will be "released soon". When is the exact date?
Answer 4: We are subject in part to the scientific peer review process, but it should be ‘officially’ released this fall. A beta version is available for both V2 and V3, if you are interested in testing, you can contact the team for installation instructions.

Question 5: Can you use remote sensing variables like Leaf Area Index and ndvi as input variables in SDM?
Answer 5: Yes, absolutely, but it is important to make sure that your occurrence data and the RS variables you are using match up in terms of temporal scale. As long as the variables are of the same extent and resolution, you can use them together. Refer to Part 1 about Leaf Area Index and NDVI within a SDM.

Question 6: How would you select a proper thinning distance? Why should the occurrence data be spatially thinned?
Answer 6: Ideally you would have some information about the biology of the species such as home range or daily travel distance, or about the sampling density to inform your selection. Another thing to consider is the resolution of your environmental variables. In your choice, try to balance to allow enough data to characterize the full environmental spaces that the species occupies while also avoiding spatial bias and autocorrelation.

Question 7: Is there a way to check the multicollinearity between the downloaded bioclim variables in wallace app?

Answer 7: In V2 you can look at a PCA in the Env Space component but as of right now Wallace does not offer a formal multicollinearity test. However, when running maxnet you can penalize complexity by increasing the regularization multiplier value; this will in practice also penalize the use of collinear variables.
Technically, it is not necessary to control for this issue if you use Maxent, however; Maxent is a machine-learning algorithm, and determines predictor variable importance in the context of other variables using internal variable selection called L1 regularization (Elith et al. 2011). The regularization algorithm can discard redundant information and gain novel information even when that novel information is relatively small. Therefore, removing highly correlated variables from the analysis could remove a small amount of data that might potentially be important. Further, collinearity is expected to have a very limited effect on the predictive ability of optimally tuned Maxent models (Elith et al. 2011, Radosavljevic and Anderson 2014).

Question 8: If you use your own data, what sort of format has to be? and how do you upload this to the Wallace?
Answer 8: There are mainly 2 types of data that you need to build an SDM a) Occurrence data, b) Predictor data. Occurrence data needs to be provided as .csv table with at least 3 columns: scientific_name, longitude, and latitude. Note that the names of the columns must match those names in that order. Predictor variables (typically environmental data) must be provided in raster format. Wallace, allows for several raster formats (.grd, .tif, ASCII) to be uploaded.
To upload the data you must go to the appropriate component and select the 'User-specified' module.

Question 9: I have one question about the modular approach of Wallace: Does Wallace simply import the functions from other packages (e.g., dismo) or does it port the functions from the source code?
Answer 9: Wallace does import functions from other packages although several have been imported into "wrapper functions" that prepare the data needed for those packages to work as well as produce the warning and error messages you see in the log box. Within each component and module in the lower left you can see a list of the used packages. In V2 this will also be provided as part of the RMD file.

Question 10: What is the basis for partitioning occurrence data? What does partition do with the occurrences? and what are the basis to choose different methods for partitioning?
Answer 10: The goal of partitioning is to split our data into groups for model calibration (building the model) and evaluation (testing how well our model predicts data that was not used to build the model. The partitioning method used will depend on your data (how many points you have) and the goal of your study. You can find more information on this topic in Muscarella et al. (2014) Methods in Ecology and Evolution.

Question 11: For the users that are somewhat new to SDM, does the Wallace documentation provide either guidance and/or literature references to all of these steps and decisions so that users can make more informed parameterization decisions?
Answer 11: One of Wallace's goals is to be instructive, providing users information and references on conceptual and methodological SDM topics. You may find guidance information in the "Component guidance" (for the general component) and "Module guidance" (for the specific tool used) tabs.

Question 12: Querying GBIF using Wallace is not working. The progress bar is stuck at 1/8 of the bar. Is this a bug?
Answer 12: I just tried it and it works fine for me - double check that you have the correct scientific name listed. You could go to the gbif.org website and search there to make sure there are data available for the species you're interested in. And double check that your "set maximum value of occurrences" is not zero - if you put a really high rumber there it could take awhile as well.
It can also be a connection issue, do try again when not streaming the webinar.
GBIF: https://www.gbif.org/

Question 13: After downloading the Wallace code, is it possible to re-run that code in Wallace? Or would you need to open the code in a text editor and use it to re-enter your model in Wallace?
Answer 13: The RMD part can be re-run inside R but not in Wallace. Starting with V2 you will be able to save and load your Wallace session as an RDS file.

Question 14: Can this model be used for plant species eg Mangrove? In modeling mangrove distribution and relating it to sea-level rise but historical sea-level rise data is

limited, are there available proxy data? Should I follow the resolution of the mangroves?
Answer 14: Absolutely yes. Theoretically it should work even better for plants because they (most of them) don't move!

We also have an ARSET mangrove training: https://appliedsciences.nasa.gov/join-mission/training/english/arset-remote-sensing-mangroves-support-un-sustainable-development

Question 15: Does Wallace have projection features for marine environments or just terrestrial? can wallace be used to model the distribution of marine species? are there any datasets related to ocean bathymetry? Most of the ocean datasets seem to be SST.
Answer 15: We are working on including marine variables for future versions. But users can upload and include any data that they want - there are marine projections available online including bathymetry yes, so you could download and then use in Wallace as user-specified variables. We will be briefly discussing marine environments in Part 3 this Thursday.

Question 16: how we can export the final SDM map to any other software such as ARCGIS for classifying area of more suitable and less or moderately suitable habitat ?
Answer 16: You can use the Download tab within the visualize module to download the rasters of your model. Several formats are available including GeoTiff, ASCII and GRD grid files. For classifying, you can apply thresholds within Wallace and download the raster of the binary map.

Question 17: Could you repeat please, what statistics in Wallace are useful to test the model quality?
Answer 17: The component and module guidance provides some detailed information about this, but some general rules of thumb are to select the optimal model based on performance metrics that balance predictive ability, model sensitivity, and model specificity, e.g. the (average test) omission error and the (average test) area under the receiver operating characteristic curve (AUC). With smaller sample sizes or for users who are concerned about overly complex or overfit models, it is also advisable to choose the model with the lowest delta Akaike Information Criterion (AICc – small sample size corrected) score, and/or by looking at the number of parameters included

in the final model. All of this information can be viewed on the table in the results tab of Wallace.

Question 18: Can you model more than one species at any given time?
Answer 18: Currently, this is not possible in V1 but multispecies modeling will be available in v2.

Question 19: As student can I Publish paper about using Wallace's data for any desired Area? if yes, is there specific guidance to do so?
Answer 19: As we state on our website, if you use Wallace in your research, please cite this paper: Kass JM, Vilela B, Aiello-Lammens ME, Muscarella R, Merow C, Anderson RP. (2018). Wallace: A flexible platform for reproducible modeling of species niches and distributions built for community expansion. Methods in Ecology and Evolution. 9:1151–1156. https://doi.org/10.1111/2041-210X.12945
However, the datasets do not belong to wallace, we just deliver them for users to run through our workflow - please view the component and module guidance in Wallace for appropriate references for related datasets and cite those in addition to citing Wallace.

Question 20: How to do modeling with Wallace on aquatic species that are limited to water areas such as rivers, lakes or swamps? and how to do modeling on migratory species that are sensitive with the barrier (like DAM) in their migratory route?
Answer 20: You can include as user-specified variables shapefiles of streams or rivers or barriers to inform or clip your model to the specific areas of interest. Clipping will be made available in v3 of Wallace.

Question 21: At Step 4 (Process Envs) I cannot find a way to select the study region. When I try to sample the background I keep getting "! ERROR : Before sampling background points, define the background extent.". Could you clarify this step? Thanks.
Answer 21: In component number 4 you have to first select the background extent either using a pre-determined extent (e.g minimum convex polygon, point buffer), a user provided extent or a user drawn extent. After creating the extent you can then sample background points within that extent.

Question 22: In Wallace you can make principal component analysis graphs?

Answer 22: In version 2 we will have a PCA in the Env Space component, so yes, soon! It will only work with two or more species.

Question 23: Are there are specifically Wallace workshops that we could participate in?
Answer 23: We have one coming up soon in October for the Student Conference on Conservation Science - New York: https://www.amnh.org/research/center-for-biodiversity-conservation/convening-and-connecting/sccs-ny
And we often have workshops at other international conferences like IBS and others, and in multiple languages. We also have recorded webinars available on our website in multiple languages. But if you're interested in hosting the wallace team or collaborators for a workshop, please reach out and let us know (wallaceEcoMod@gmail.com).

Question 24: How long should Wallace take to spatially thin user input occurrence records? I've noticed that for 2k records it takes my computer an hour. When I tried running several thousand it wasn't done in over 9 hours.
Answer 24: Wallace uses the spThin package for thinning. Have you tried thinning the data directly in R using the spThin package? The speed will really depend on that underlying package and your own computer power. What spThin is doing is measuring the distance between your point and other points. More points means more comparisons and more time.

Question 25: Thank you for the presentation! In the past I had problems using wallace with species that have broad distributions or a lot of occurrences, Wallace usually fail to build the polygon, or do the SDM, is this something that will change in the new version?
Answer 25: The new version should prevent many of the grey screens. However if this problem is about computer power it might persist. Version V2 does have an option to run models in parallel which should help address this problem. If it is at the polygon level, you may need higher computing power.

Question 26: Any rule of thumb minimum occurrence data that can be used in SDM? especially MaxEnt to get good habitat prediction. Would you still recommend spatial filtering as a way to account for sampling bias when you have few occurrence records (e.g. less than 20)? Is there any other way of accounting for sampling bias that doesn't run the risk of reducing already sparse occurrence records?

Answer 26: Good question. Generally, >30 occurrence points is considered enough to build a robust model. However, “decent” models can be built with as few as 10 points. This is largely dependent on how large your modeling extent is and how much environmental variation there is within this area. In terms of spatial filtering, it depends on how clustered your points are within that area. It takes a little bit of thinking and background knowledge of your area (also see the answer to Question 6).

Question 27: We have been working on crop plagues caused by different insects...is Wallace suitable to project expansion of insects as well as fungi and bacteria knowing their occurrences?
Answer 27: Yes, absolutely as long as you have the occurrence data for these species, which may not be available in online databases (e.g. GBIF, VertNet, etc.). However, if biotic interactions are expected to play a major role in the distribution of your insects/bacteria you may want to consider methods to include these into your SDM workflow. Consider reading Anderson (2017) “When and how should biotic interactions be considered in models of species niches and distributions?” J. of Biogeography.

Question 28: What are assumptions to take into consideration when using camera traps occurrences points for SDM?
Answer 28: Mostly sampling bias - carefully controlling the background training extent and spatial thinning of points might be quite important for this type of data.

Question 29: can the model be used in crops such as rice? Or be used to predict microorganisms distribution?
Answer 29: In theory, yes. However, you must carefully consider what are the drivers of your species’ distribution.

Question 30: Ff I have an archive of occurrence data of different species that was taken from november/2017 to december/2020, what date should be the best for the environmental variables if I want to upload to Wallace?
Answer 30: Climate variables are 30-yr averages so you should be good there, but if you’re talking about remote sensing data, perhaps choose a median date/year, but double check that there aren’t any drastic changes at occurrence points e.g. recent forest loss between the year of the point and the year of the RS data you will use. In R, you can also make a point process model, where you can make environmental

variables for each point. Samples With Data (SWD) in Maxent is also capable of this as well. We are thinking about incorporating this into Wallace.

Question 31: If we do not have point occurrence data of species, could we do SDM like study using IUCN distribution data, which is in polygon form?
Answer 31: Strongly advise against! But people do do it sometimes by sampling points randomly within the polygon. In part this depends on the resolution of environmental variables included.

Question 32: Can wallace be used to model distributions/predict for highly mobile species especially migratory species of birds and fishes? This is particularly for species that have very poor ecological/biological information.
Answer 32: With extreme caution and with thoughtful inclusion on resolution of environmental variables or perhaps with some creative environmental variables that capture the migration process. It would be important to do separate models for diff seasons e.g. only for breeding season. In Part 3, we will be discussing another model called Circuitscape that you can also use for movement patterns.

Question 33: Is it possible to say that Wallace is a simplified and advanced form of MIAmaxent (A Modular, Integrated Approach to Maximum Entropy Distribution Modeling)?
Answer 33: We're not very familiar with it, it seems like a modelling method to "replace" Maxent. Wallace is expandable and can include many/any modelling technique. The advantage to Wallace is that it helps with best practices in data pre-processing and model parametrization.

Question 34: It is possible to visualize/save response curves?
Answer 34: Yes! In the Visualize component you can look at all response curves. Using the Download tab in that component you can also download them.