Part 1 Question & Answer Session A

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 Justin Fain (justin.j.fain@nasa.gov) or Sativa Cruz (sativa.cruz@nasa.gov).

1. **Question: How can I ensure that true absences are true, especially for plant species which do not move around? Sometimes the absence is just because we don't do enough surveys or simply the species is overlooked, like when there is a lack of flowers so a species cannot be identified properly.**
    a. This is the detection probability problem. If you have enough data, you may be able to use hierarchical models and account for detection probability, correcting for false absences. To do so, you need to conduct repeated surveys. Check the occupancy modeling literature, for instance. For the models we are discussing, you cannot know if absences are true absences or not. Thus, we refer to these points as pseudo-absences.

2. **Question: Is there a workaround for determining absence data on narrowly endemic trees?**
    a. Sometimes you may be able to infer absence from surveys. But this relates to the previous question on detection probability. See answer to Q1.

3. **Question: Do we always need a presence and absence dataset? It is sometimes difficult to have a true absence of a species. Could a presence dataset be sufficient, for example one from an herbarium?**
    a. Having presence and absence is the best-case scenario. But we recognize that often, absence data are not available. We will describe incorporating pseudo-absence data in our hands-on workflow.
    b. Using data from an herbarium is a valid strategy. You just need to carefully think about predictor variables that align to the type of presence data you have.

4. **Question: Since Part 1 covers pseudo-absence integration, what is the best practice within GEE to define the spatial boundaries or environmental profiling for generating pseudo-absences? How can we prevent cloud-**

based random sampling from selecting points in areas with high historical sampling bias, like near roads or urban herbaria?

a. We will cover the different approaches available in this workflow to generate pseudo-absences in Part 2, illustrating an environmental profiling technique. In doing so, you can mask areas you do not want pseudo-absences to be created.

5. **Question: If I have very few occurrence points (like 66) and a very small area of training based on which I will be projecting to a bigger area, I have observed maxent projections are quite wrong and it learns very easily from the presence point features. How do we handle those?**
    a. There is no solution to lack of good, quality data. Unfortunately, the only solution is to increase the sampling effort and collect more data. You should be very careful in predicting areas that lie outside the range of values you used to fit your models. We will talk more about this in Part 2.

6. **Question: How can observed species movement patterns be used as early warning signals for detecting ecological changes caused by climate change, land-use change, and watershed environmental stress?**
    a. Doing so requires a workflow that is different from the models we are presenting here. The general concepts would be the same as what we are discussing with SDMs, but since movement data are autocorrelated in space and time, other models like integrated step selection functions (iSSFs) are more appropriate.

7. **Question: Is there a way to include environmental history as a covariate (e.g., if the soil type changed over the last 5 years) when using SDMs?**
    a. Yes, you can match the time the presence data was collected with the time of the covariable. We will not cover how to do that in this workshop, but if you look at Case Study 2 in the original paper, you can find the code, which should be easier to understand after completing this workshop.

8. **Question: How can observed species movement patterns be integrated with species distribution models, remote sensing, and hydrological/environmental predictors to estimate changes in habitat suitability, ecological connectivity, and future species persistence under climate and land-use change?**

   a. That is a good question and probably goes into very advanced models. There are ways to integrate datasets using hierarchical models and Bayesian approaches. A simpler approach would be to use GPS locations from movement data as your occurrence data in an SDM. You can see an example of that with the code listed here. With species movement patterns, you need to be conscious that the points are autocorrelated in space and time (i.e., a point collected at time *t* is highly related to a point collected at time *t-1*, especially when data are collected at a high temporal interval (<1 hour). To properly account for the autocorrelation that exists, you'd need to either thin your data until the points are independent (a key assumption of an SDM) or conduct an integrated step selection analysis. The concepts of an iSSF are like a SDM but developed to accommodate the autocorrelated structure of your data.

9. **Question: How can species distribution models, remote sensing data, and movement observations be combined to develop a dynamic ecosystem early warning system?**
   a. See the answer to the previous question as well as this ARSET training on Animal Tracking from last May.

10. **Question: Since GBIF datasets require heavy filtering for quality control (e.g., removing geospatial issues, spatial duplicates, and old historical records), is it better to perform this filtering directly on the GBIF portal before downloading, or does GEE have efficient workflows to filter the raw GBIF cloud asset programmatically?**
    a. I will advise doing all the filtering before uploading the data to GEE as an asset. GEE is much better at handling raster data than vector data. You can do it, but it will consume memory and slow things down. You can do the filtering in the GBIF portal or using R or QGIS for instance.

11. **Question: Is there any workflow, we can import data from GBIF using API?**
    a. I believe it is possible to do with the Python library, but I am not aware of a way to do it within the GEE code editor.

12. **Question: For the community GEE catalog, do the datasets provide instructions for cleaning and quality checking, or is that already completed?**

a. All metadata information is provided for each dataset in the portal and is generally associated with a peer-reviewed publication.

13. **Question: How about marine species distribution? Are the methods like the terrestrial ones?**
    a. Yes, you can apply the same techniques to marine organisms. There are many datasets available to use as predictor variables for the seascape.

14. **Question: How do you incorporate the temporal scale into Machine Learning algorithms? If you have records from 2000 through 2020, how Maxent is using the year-matched environmental predictors information?**
    a. See question 7 -> Yes, you can match the time the presence data was collected with the time of the covariable. We will not cover how to do that in this workshop, but if you look at Case study 2 in the original paper, you can find the code, which should be easier to understand after completing this workshop.

15. **Question: If my primary objective shifts from predictive accuracy to estimating parameters and evaluating variable relationships (inference), which alternative modeling techniques should I consider?**
    a. You should consider using a parametric approach like Generalized Linear Models or Generalized Additive Models. To do so, true absence data is required. Even better, use a hierarchical model that accounts for detection probability if you have the correct type of data.

16. **Question: When conducting environmental profiling for pseudo-absence generation, what kind of parameters are used to conduct dissimilarity tests? If the same variables or raw data is used, for example, using optical and topographical data for profiling and then using similar data for SDM, does not that artificially improve the classification? Is an inherent bias introduced, inflating the model performance evaluation metrics?**
    a. A very good, important point. Different types of datasets (presence data) come with different challenges. It is possible you are artificially improving the model, which may work for some species. That is why in Part 2 we emphasize the importance of having independent data to validate SDM predictions and not solely rely on internal validations (i.e., using the same data for validation that you are using for fitting the model).

17. **Question: The previous speaker mentioned many unobserved confounding factors, such as predation, competition, etc.… What counterfactual sampling strategy (absence, pseudo-absence and background data) and algorithm are best suited to deal with so much uncertainty and why?**
    a. In any model, there are assumptions you are making. Incorporating biological effects, such as predation or competition, is challenging. Generally, you must acknowledge the limitations of the model. There are modeling strategies to incorporate the distribution of another species as a predictor variable for your species of interest. How good a species distribution of a predator can adjust for the habitat suitability of a prey is also debatable. But that is related to the predictor variables and not the pseudo-absence data generation process.
    b. Machine learning models tend to be very good at handling noise and performing well. But how well models predict can depend a lot on the data you have. That is why companies are so interested in gathering so much data to improve model predictions and why data today is so valuable.
    c. Another similar assumption when doing climate change predictions is assuming species do not adapt and in 50 years the species will continue responding in the same manner to the predictor variables, which is likely not true. Those are limitations. All models are wrong, some are useful.

18. **Question: Not all species have the same tolerance thresholds for ecological and/or climatic variables. How should the generalization of pseudo-absences take this into account? What criteria determine whether the model is reliable?**
    a. Good question. We will cover model validation in Part 2. We will discuss different metrics to assess model accuracy.

19. **Question: Is there a function that can be used to quickly clip to "land" or "not land"?**
    a. Yes, that is called masking. Grant is explaining that right now. We will show how to use a mask in Part 2 in the SDM workflow to mask out “not land.”

20. **Question: If we want to do a particular thing, we write a code to do that thing through computers, but what if we don't know that language? Can AI be used to bypass the language learning curve?**

   a. AI can certainly be used to decrease the learning curve when conducting an analysis, especially related to learning new coding languages. I would not say AI allows you to completely bypass the language learning curve but can be very helpful. Importantly, you need to be specific with what you are asking AI to do for you (i.e., How do I clip an ee. Image with a polygon or ee.Feature in Google Earth Engine (JavaScript or Python)?). The more concise you are with your question, the better your results will be. AI will then help you with proper syntax that you can build upon.

21. **Question: In generating pseudo-absences, is there a standard number of these pseudo-absences and their distance threshold to presence data? I think I would have to change some of these settings if I were to use the same method for a smaller study site (e.g. small islands).**
    a. Good question. We will cover this in Part 2. In machine learning methods like random forest, a best practice is to use a balanced dataset (i.e., same number of presences as absence).
    b. The distance threshold to create pseudoabsences can be quite arbitrary. It is a good practice to have a supported argument to determine that distance. For instance, you can use the maximum distance a species can disperse.

22. **Question: Is the default spatial reference 4326 or wgs84?**
    a. Yes. That is correct. EPSG:4326 is the official identifier for WGS84.

23. **Question: I just figured out how great the GEE data catalogue is. How can I contribute to it? Locations and status of remote areas may not be up to date (protected forest, urban forest, ...). How do I ensure that data are frequently up-to-date and contribute to it?**
    a. There are various remotely sensed datasets that have yet to be incorporated in GEE. Slowly these datasets are being integrated, but they do take time. The more the community is asking for a particular remotely sensed layer impacts how these layers are prioritized for incorporation in the catalogue. For specific information about your study area (e.g., an update to your study area boundary that does not exist in the catalogue), you can upload these into the code editor as an asset.

24. **Question: When using a bioclimatic variable, how to handle occurrences that extend before and after the 1980 - 2000 window when analysis needs to keep that occ. data?**
    a. Unfortunately, you are limited with the data you have access to form a statistical relationship.
    b. Some people might filter to keep only occurrence data from the time period matching your climate predictors. However, if the change over time in climate is very small compared to the spatial variation in climate conditions over the species distribution, then it probably doesn't matter much to have a temporal mismatch between climate predictors and occurrence data.

25. **Question: Can we use AI agents to automate species distribution models using GEE?**
    a. Refer to Question 20. AI agents can help with your syntax and are a great resource to reduce the learning curve but are not substitutes for learning to code.

26. **Does the Python language work in google earth engine or do we need to use Java?**
    a. Earth Engine supports both JavaScript (in the web-based Code Editor) and Python (for use in notebooks like Colab). While the syntax differs slightly, both approaches perform the exact same spatial analysis on Google's servers.

27. **How important is collinearity diagnostics in SDM?**
    a. This is a good question. It depends on the modeling framework you are using. Machine learning techniques are fairly robust to collinearity, though it can affect variable importance metrics. But a GLM can have more issues, particularly when your goal is to make inference (i.e., understand the drivers of species habitat suitability).

28. **Is there a set catalog or is that reading from a local working directory?**
    a. Not sure to which dataset this is referring to. Google Earth Engine has a catalog from which you can read lots of images, image collections, features, and feature collections. But some data may be available only by users, like the occurrence data we will use in the model for part 2. If there are other datasets (e.g., a local soil layer) that you'd like to use in

your workflow then you would need to import those datasets to GEE as assets.

29. **Could we calculate principal components using GEE? Can we export as different layers georeferenced to be included in QGIS?**
    a. We have not worked with principal components analysis in GEE but did just find that a custom function has been written for this. https://developers.google.com/earth-engine/guides/arrays_eigen_analysis
    b. Data layers can certainly be exported in GEE so that they can be included in QGIS. The first lecture today showed a few examples of code for how to export results to Google Drive.

Part 1 Question & Answer Session B

Please type your questions in the Slido/Question & Answer Box. We will try our best to get to all your questions. If we don't, feel free to email Justin Fain (justin.j.fain@nasa.gov) or Sativa Cruz (sativa.cruz@nasa.gov).

1. **In the slide "Interpreting Species Distribution Models," could you clarify what 'true absence data' means methodologically? How is it collected (survey design, sample size, spatial coverage), and does the protocol account for seasonal variation, species activity patterns, or unusual events (fire, floods) that could affect detectability?**
    a. Hopefully, this was answered in the lecture. In the context of this slide, we intend 'true absence' to mean data derived from robust field surveys where the species was actively searched for but not detected, distinguishing it from pseudo-absences or background points which are generated without empirical field validation. There are methods for accounting for false negatives through repeated surveys to estimate detectability, which can factor in season, activity patterns, and other factors. To emphasize, if presence and absence data are available, use it. If you have data to account for detection probability, even better. Seasonal variation can be accounted for by modelling different seasons separately or including seasonal variations in the predictor variables (e.g., standard deviation of vegetation productivity).
    b. There are other models (e.g., dynamic occupancy models) that can deal with colonization and extinction in sites and changes across time, but not in the modeling framework we are presenting here.

2. **I am wondering about fireflies currently. Are the models weighted to key variables? How is the niche characterized effectively?**
    a. In the Random Forest models we are using, variable importance is determined algorithmically; the model objectively weights predictors based on their power to explain the species' distribution (the outcome of the model). The strength of the model at predicting is assessed using validation metrics. We will cover this in Part 2.

   b. That said, which predictor variables you use in the model is defined by the researcher. That is when knowing the ecology of the species and the question you are interested in is critical. You really want to make sure to use biologically meaningful predictor variables, or the SDM model will do a poor job representing the distribution of your species.

3. **Similar question to question 2: I've made a field app for firefly occurrences. Any tips for data collection that would make it easier for data transfer in GEE?**
   a. You can upload csv files or shapefiles as assets to Google Earth Engine (GEE). We will see how to upload a csv file as an asset in Part 2.

4. **How do you connect the images of all the occurrences?**
   a. This will be shown in Part 2. Basically, we extract the pixel values at the location of each point. Thus, the spatial resolution you choose for the model is important. More about this next week.

5. **Does SDM GEE model use rasters of different resolutions, because I used Maxent and it requires ascii files same columns and rows?**
   a. Yes, Google Earth Engine (GEE) can handle rasters of different resolutions automatically, unlike the traditional MaxEnt software which typically requires all input files to be pre-processed to the same grid (extent, cell size, and projection). GEE does “as needed” computation and basically resamples as needed to give you the output you requested (e.g., an SDM output on a certain grid).

6. **Is the practice of using an ensemble approach (weighted predictions from several algorithms) less commonly utilized now in favor of focusing more prominently on machine learning methods like Random Forest?**
   a. There is a lot of debate about this. Some research shows that ensembles seldom outperform the best individual model. Others argue that averaging across model outputs can provide more generalizable model predictions or may be more robust to overfitting. We will discuss this more in Part 2.

7. **Similar question to Question 5: So, you can't generate the type of variable importance plots in GEE that you can in the Random Forest package in R (mean decrease in accuracy, gini)?**

   a. You can extract the variable importance. We will show this in Part 2 but will also discuss some limitations in terms of how they are interpreted (i.e., they are not). You can export the values and plot them in R or other software of your choice. You can also generate simple plots in GEE, but it can be much easier to do in R or Microsoft Excel/Google sheets.

8. **How accurate are the images from Landsat? From now, what is the lag in days/months of actual GE images?**
   a. Landsat images have very high spatial and radiometric accuracy. GEE offers different Landsat product “Tiers”, including Real Time and more processed datasets. I believe the latency for Real Time Landsat imagery is a couple of days - so they appear in the GEE catalog very shortly after being acquired by the satellite.

9. **Can we use SDMs for microorganisms, such as pathogens?**
   a. That is a great question. You can, but for pathogens that rely on a host, you may be better modeling the host habitat suitability or using the host habitat suitability as a predictor variable.

10. **Can I apply this with my own parameters? I work on benthic species and therefore curate my own environmental variables from available metadata. Would I be able to plug in my own data in the program?**
    a. Yes, you can upload your rasters and shapefiles as assets to GEE. In part 2, we will show how to upload an asset.

11. **What happens when a species is known to be in very shallow waters, but you have an occurrence very deep, which is something very weird; how could we show it in a prediction?**
    a. If a species strictly inhabits shallow water, an occurrence record in deep water could be an outlier that you might want to try to detect and filter out before modeling. If the bulk of your data are from shallow water, however, the most likely situation is that the deep water will show a lower habitat suitability than shallow water in your outputs despite a single outlier.

12. **How do you clean the data according to your specific area of interest when data downloaded from GBIF fall outside?**
    a. You can use filters in GEE like filterBounds() to retain only points within your area of interest. You can also first open the occurrence data in QGIS

or similar software and remove records that are clearly wrong (points on a zoo as shown in the video).

13. **Can you provide the GEE JavaScript of the model please?**
    a. It is all available in the manuscript and in the PDF of the class.

14. **Is there a pipeline for actionable information outputs in any country that aligns with species at risk legislation?**
    a. I (RDC) am not sure about this. It may exist in certain countries. The IUCN website has distribution maps for multiple species. The goal here though is to map species habitat suitability - which can certainly be useful for species management or regulatory decision-making.
    b. The Group on Earth Observation Biodiversity Observation Network (GEO BON) is trying to make data, such as repositories like GBIF, more accessible by providing reliable outputs to policymakers and government officials. We are working on this with animal movement data, with a paper that was just recently published of this urgent need (https://geobon.org/new-paper-calls-for-animal-movement-indicators-in-biodiversity-policy/) and there is information from Bon in a Box (https://boninabox.geobon.org/indicator?i=SDM) for species distribution models. I (JAS) believe there is still considerable work that is needed to make the connection between quantitative outputs and the policymakers that want/need this information to make more informed decisions.

15. **Can I use an R script in GEE?**
    a. There is an R package called rGEE that allows you to run GEE in R (using very similar function names). But you cannot use R code in the code editor of GEE. They use different programming languages.

16. **Can we store a set of datasets (i.e. species coordinates in a dictionary)?**
    a. Your dataset will be a feature collection. In a feature collection, each point has a coordinate and properties attached to it, like whether the point is a presence or an absence, and the pixel values of predictor variables.

17. **Can one calculate the spread debt of the species in Earth Engine from the prediction maps constructed?**

a. Certainly. You can do all kinds of image calculations in GEE. You can also export the model outputs and perform those operations in another software if that is easier for you.

18. **Is there a rule of thumb to decide between solving simple operations in our console (i.e. concatenate simplestring+simplestring) vs passing the operation to the EE server?**
    a. This is something you will get a feel for over time. If you are just typing basic text or doing simple math for your own notes, you can use regular JavaScript code. However, the moment you need to use some object or value inside an Earth Engine function you will generally have to wrap it in an ee constructor (like ee.Number() or ee.String()) so it can be used on the EE server.

19. **When will we need to define a subject in EE language versus JavaScript?**
    a. This is like the last question. You need to use a “container” (the ee.thing) anytime you are requesting data processing on the Google server and not on the client side (your browser). For instance, if you want to divide the NDVI value of a MODIS image by 10000 to scale it to the -1 to 1 scale, the image must be an ee.Image.

20. **Does python language work in google earth engine or do we need to use Java?**
    a. Earth Engine supports both JavaScript (in the web-based Code Editor) and Python (for use in notebooks like Colab), and while the syntax differs slightly, both approaches perform the exact same spatial analysis on Google's servers.

21. **Can this methodology be applied to insects in tropical agroecosystems?**
    a. Certainly. The methodology can in theory be applied to most ecosystems on Earth, sea or land. However, each study needs to be carefully planned in terms of the spatial resolution, temporal resolution (if applicable) and extent.