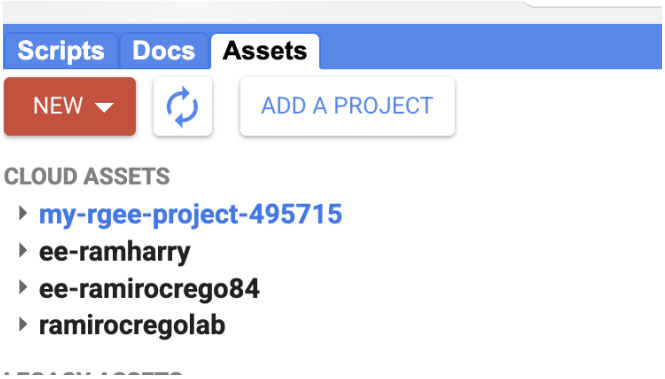
Part 2 Question & Answer Session A

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

1. **Question: When you download data from GBIF, what is the analysis area extent? Can I do it for a location radius 10km for instance? Can I analyze a much larger area? Also, does it work in the whole world?**
    a. You can set the extent in GBIF (for instance, restrict data to a bounding box or a specific country). You can certainly download the data for the entire world too. You can also select points in GEE with a filter. There is a lot of flexibility in getting the dataset you actually need in your modeling exercise.

2. **Question: When uploading data, where is it saved? Is it connected to my Google account or is the Earth Engine account separate? Thinking about safety for coordinates that should be kept secret for protected species and such.**
    a. The data is saved as an asset, and it takes storage from your Google account. It is all connected with your Google cloud. You can see all your assets in the asset tab.



In terms of sharing, the data will be accessible only to you unless you specify either a specific person that you want to allow access or by making the data publicly available. You can check again in the recording where I showed where to click to make the data public.

3. **Question: When defining spatial resolution, does GEE automatically upscale or downscale the collected data, or must a custom script be written to process each dataset manually?**
    a. Google Earth Engine by default operates on a "pull" basis, where it resamples your data (and any other data you are working with from the catalog, Awesome GEE Community datasets, etc.…) only at the point that you request an output at a certain resolution. So, it may be only at the export step that you define a specific spatial resolution (e.g., 30m) and GEE will handle the resampling and reprojection automatically during the export process. You don't need to be constantly upscaling, downscaling, or resampling every time you bring in a new dataset in your script.

    One thing to note is that when you add outputs to the map viewer window in GEE, the resolution of the output will depend on your zoom level. You can force GEE to first resample to a given resolution or grid, but typically you would use the map viewer just to explore/problem solve and then export your final outputs at your desired resolution as a last step.

    Another important note is that it is good to familiarize yourself with how GEE handles this on-the-fly or final resampling. For instance, the default is that GEE uses nearest neighbor resampling and this may not be what you want.

4. **Question: In the case of working with parasitic organisms, what is the best strategy to include the host in the variable? Is it by doing the SDM to predict the distribution of the host and then feed the results on the next SDM for the parasites as a variable?**
    a. Your proposed approach of first modeling the host and then modeling the parasite using the host distribution model as a predictor is a valid one. If you can and have the data, you could do a hierarchical model in which the presence of a parasite depends on the host being present. But that requires a completely different modeling framework.

5. **Question: Why do we delineate a buffer?**
    a. It is one approach for defining your area of interest. But it is not the only approach you can use. In this case, we are defining the area of interest

around a buffer to all presence data. The buffer ensures no presence point is at the edge of the area of interest. Also, note that the area of interest used to fit the model can be different from the area of interest used for predicting. For training your model, you want to have an environmental space (area of interest) that is representative of your species of study.

6. **Question: What are these points shown? And what is the accuracy of points? GPS?**
    a. Displaying the points can be very useful to have an idea of the spatial distribution of your data. For instance, a very clustered dataset will be difficult to model. It can also be important to visually check that presence points are beings masked out from the area to create pseudoabsences. The accuracy of the points depends on your data. We showed in Part 1 how you can restrict the uncertainty you want to tolerate in your data in GBIF, which should be linked to the spatial resolution of your model.

7. **Question: How can I use NDVI in prediction of species potential suitable habitat for future climate scenarios? Is that a static variable?**
    a. NDVI is typically calculated from satellite imagery, so it does change over time (i.e., it is not inherently a static variable). However, I don't know of any datasets that predict future NDVI. So, if your SDM includes both NDVI and climate variables then your future suitable habitat predictions would likely have an implicit assumption that you are showing the predicted change in habitat suitability due to climate change under a scenario where NDVI does not change.

8. **Question: What about using the AlphaEarth Embeddings now too?**
    a. AlphaEarth Embeddings are an asset on Google Earth Engine. This means that they can be integrated into your analysis like any other ee.ImageCollection. A full description of the asset is available at the link provided.

9. **Question: How about the limitation (or not) using presence/absence ratio. Can it be 1:1 for ensemble modelling (irrespective of maxent/random forest specific)?**
    a. Yes, in this example using Random Forest, we are generating balanced datasets with a similar number of presence and pseudoabsences,

following results from Barbet-Massin et al. 2012 (DOI: 10.1111/j.2041-210X.2011.00172.x) who showed that balanced datasets give better performance in machine learning algorithms. But you can modify this and use different numbers of pseudoabsences to presence.

10. **Question: I didn't quite understand the cloud mask from the previous session. Is this to remove satellite imagery with too many clouds to see?**
    a. There are many ways to handle clouds when working with satellite imagery in GEE and there are lots of good examples of code for this type of pre-processing with different satellite datasets in GEE. The basic example in the last session showed a process of 1) first filtering your image collection to include only the least cloudy images (e.g., dropping images with >10% cloud cover) and 2) then creating a composite image by taking the median pixel value for every pixel in those filtered images. This gets you a cloud free mosaic image to work with in most cases. A more robust option is to add one additional step: 1) filter to less cloudy images as before, 2) *mask out cloudy pixels from each image using a custom function that references an image QA band*, and 3) then create your composite image by calculating the median or mean of all cloud-free pixel values.

11. **Question: How do you decide which 6 bands to keep for the habitat suitability model?**
    a. That is when understanding the ecology of your species of interest is critical. If you are working with Landsat imagery, it is common to use the visual, near infrared, and shortwave infrared bands as these typically vary among different vegetation types/conditions. Sentinel-2 also has “Red Edge” data that could be used. Generally, however, you should select a set of predictor variables that makes biological sense, that you expect to be drivers of the distribution of the species. For instance, if you are modeling the spatial distribution of a plant, then soil characteristics like percentage of sand in the soil, would be a good predictor variable.

12. **Question: Could you run some sort of collinearity study (like VIF) to eliminate highly collinear variables in JavaScript before modelling?**
    a. Yes. Some SDM models are fairly robust to collinearity (e.g., random forest) while others are not (e.g., logistic regression). However, it is often

a good idea to check for highly correlated covariates as a preliminary step. It is also worth noting that variable importance metrics can be influenced by collinearity since having a bunch of correlated variables in a model tends to reduce important statistics for these variables.

In the original code we published, there is code to check the correlation among predictor variables.

13. **Question: Ultimately, SDM would not be realistic without accounting for human behavioral data. Do you have some snippets that account for data on human behavior, such as hunting, cattle grazing, and infrastructure, such as roads?**
    a. This is very valid point. A commonly used product is the human modification index, which accounts for roads, land cover, livestock, etc. (dataset).

    Best is to look through the Data Catalog or the Awesome GEE Community datasets to determine what layers are currently available in GEE. For instance, you can use raw road data in GEE (e.g., TIGER Roads data for the US or the Global Roads Inventory Project (GRIP) dataset from the Awesome GEE Community datasets) to calculate road density or even import your own datasets into GEE to meet your project objectives.

14. **Question: My presence points are concentrated in a small space. Even after spatial thinking, I had around 100 points out of 115 points obtained via the GBIF portal and my predictors included OSM layers where I had proximity to roads, places, and powerlines as they affect the particular species along with NDVI and LST. When I run maxent with all proper settings, it learns too bad and predicts only the western side where the presence points are located. What do you suggest?**
    a. Having very clustered data is a common problem with rare species. It leads to models that overfit the data, as it seems is the problem you are describing. Unfortunately, there is no magic solution to lack of data and ideally, you can collect more data from areas that are not represented in the dataset.

15. **Question: Is there a recommended ratio between background points and presence points in species distribution modeling? Does the scale of the study area affect this ratio?**
    a. See answer to question 9. This paper is very good for that: Barbet-Massin et al. 2012 (DOI: 10.1111/j.2041-210X.2011.00172.x). Different methods respond differently to the number of pseudoabsences you create. Here we use a balanced dataset as it is often recommended to achieve good performance with machine learning methods.

16. **Question: If my species of interest is endemic to a very small area such as 15sq km., can we apply SDM for the species?**
    a. You can, yes, but you will be limited by the spatial data that is available for analysis and how well your predictor variables can explain the distribution of the species. The example that you provide would need to be conducted at a high resolution, potentially requiring you to import spatial data layers for analysis to meet your project objectives. In this case, you will need to decide if Google Earth Engine is the optimal modeling platform for your analysis, as you might be able to perform the analysis locally on your machine. The workflow we describe, however, could still be useful.

17. **Question: Some locations where the species was present in the past may no longer be suitable today. So, how old should occurrence records be for habitat suitability modeling?**
    a. This is very good point. You should try to link your predictor variables to the time span of your presence data. But also note that in GEE you can link predictor variables to the time the data was collected. See case study 2 in the original paper (we briefly explained that in Part 1) in which we link the predictor variable to the year in which the presence data was collected. These time series analyses are very powerful.

       I think these types of cases also require careful model/map interpretation. For example, you may be mapping suitable habitat that is no longer occupied for some reason (e.g., human impacts) and may not accurately represent the present-day distribution.

18. **Question: Can we do python coding to pull the data from GEE without ever going to GEE?**

a. Yes. The entire code was translated to Python as well.

19. **Question: GEE facilitates large-scale analysis, but its workflows often rely on temporal composites, collapsing pixel phenology into summary statistics (median). Given the 'time-first, space-later' paradigm in the site's R package, how could incorporating pixel phenology improve Species Distribution Models (SDMs)? Moreover, are the methods presented in this ARSET compatible with using full time-series information as predictors, or do they inherently require temporally aggregated variables?**

    a. The SDM workflow presented does not inherently require certain types of predictor variables. You need three things: 1) occurrence data, 2) predictor variables as raster datasets, and 3) a model/algorithm for relating 1 and 2. So, your predictors can be anything you feel is biologically meaningful and you could do something like calculate the standard deviation in vegetation metrics over time and use those pixel values as a measure of variability in phenology. There are many published studies where people extract phenological metrics in GEE and any of these metrics can be used as predictors if they are in raster format.

20. **Question: Does the timeline for worlclim variables that are being used need to match the timeline of my presence points? Suppose my study is from the year 2020 and after. I can't use the historically averaged worldclim data, right?**
    a. This is a good point. It will largely depend on your species of interest. For some species, it may be okay to use the bioclim variables. But in GEE you can also use other climatic datasets that allow you to extract climatic variables as detailed as to the hour. See the ERA5 datasets (https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_DAILY). Note that while you gain in temporal resolution, you lose spatial resolution (e.g., 11 km), but there are many new methods coming out that allow you to downscale those types of data.

    b. You may also be more or less concerned about this temporal mismatch between your occurrence data and the worldclim timeline depending on your objectives. If you are modeling projected climate shifts, then this mismatch may be important. However, if your goal is simply to map the

range of a widely distributed species, historical bioclimatic variables are often perfectly acceptable, as the slight difference between past and present climate is usually just one minor source of uncertainty. There could even be 'lag effects' where historical climate may predict a species' distribution better than current climate.

21. **Question: How was the binary map thresholded?**
    a. This was hopefully responded to in the video. In this example we are using two approaches: the map shown is the result of the mode among the 5 iterations of the classification provided by the random forest. A more adequate approach is to use the threshold that maximizes the sum of specificity and specificity. Doing this is memory intensive so we only provide the code to export the resulting map in batch mode.

22. **Question: Is there a way to scale this up to run multiple species at once?**
    a. Unfortunately, not at the moment. But I have seen papers in which they set up the code in Python and have a function that iterates over multiple species.

23. **Question: How can we determine the optimal number of trees for a Random Forest model in species distribution modelling?**
    a. You can run different models with different specified number of trees and compare the AUC values. Then you select the one that predicts better. We are working on a code using R and rGEE that would allow users to do fine tuning of random forest specifications.

24. **Question: What's a good AUC ROC score to make the model confident?**
    a. 0.5 is not better than random. People consider 0.7-0.8 fair, 0.8 - 0.9 good, and > 0.9 excellent.

25. **Question: Could you recommend papers about using AUC-ROC and/or AUC-PR?**
    a. We have a good discussion about that in the original paper with references. Here is that paragraph:

"Despite AUC-ROC being broadly applied as a model validation criterion in SDM analyses, it is important to note that the use of AUC-ROC has been criticized as an accuracy metric when no true absence locations are available

(Lobo et al., 2008). The AUC-ROC in- creases when the area to select pseudo-absences increases outside the environmental domain of presence locations, resulting in inflated AUC-ROC scores (Jiménez-Valverde, 2012; Lobo et al., 2008). This also implies that the AUC-ROC is inflated when species of interest are narrowly distributed in relation to the extent of the study area (Lobo et al., 2008). In contrast, AUC-PR is not influenced by the number of absences (Sofaer et al., 2019). Therefore, for presence-only data, AUC-PR is preferred as it is more sensitive to the ability of the model to correctly predict presence locations, especially across large spatial extents or when modelling the distribution of rare species (Sofaer et al., 2019).”

Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: A misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography, 17(2), 145–151. https://doi. org/10.1111/j.1466-8238.2007.00358.x

Sofaer, H. R., Hoeting, J. A., & Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. Methods in Ecology and Evolution, 10(4), 565–577. https://doi.org/10.1111/2041-210X.13140

Jiménez-Valverde, A. (2012). Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Global Ecology and Biogeography, 21(4), 498–507. https://doi.
org/10.1111/j.1466-8238.2011.00683.x

26. **Question: How do we use SDM using multiple models in GEE? If our AUC is around 1 like 0.98 even using spatial block CV, how do we know if it is not overfitted and how do we minimize it?**
    a. Such a high AUC is often a red flag that you could have an overfitted model. However, it is important to note that your choice of study area and background points can influence accuracy metrics. For example, if your study area includes large, "easy-to-distinguish" true-negative areas (places that are clearly unsuitable) then this can inflate AUC scores. Your AUC scores may not be ‘wrong’ or ‘artificially high’ in this case - you just gave your model an easy test.

b. Sometimes, especially if you have a very small number of presence points, your model can include one or more predictor variables that perfectly separate presence and absence data. The model may not have learned a true ecological relationship but may have a 'shortcut' available where it can perfectly explain the presence points using one of your predictors. There's no easy solution to this overfitting but ensuring that you have a decent sample of truly independent validation data would be best practice.

27. **Question: How do you justify the use of GBIF occurrence data in Species Distribution Models (SDMs), given that they may contain taxonomic misidentifications, georeferencing errors, and sampling bias?**
    a. If you have a more reliable dataset available (e.g., trained experts conduct systematic or random survey locations), then this would often (not always) be preferable for modeling. GBIF aggregates data from many sources, including citizen science datasets, so there is definitely risk of error or sampling bias. However, GBIF data may be the only or best dataset available. The challenge is then to do the best you can with an imperfect dataset. There are lots of steps people take to clean or quality check GBIF data. For instance, you can filter for records with low coordinate uncertainty, remove obvious outliers, and apply spatial thinning to minimize the effects of sampling intensity bias, or use target group background selection to deal with spatial bias (this is often done with eBird data for instance). Ultimately, however, a few errors in your occurrence data are unlikely to dramatically affect your model outputs if you have enough clean/accurate data.
    b. On the plus side, data aggregators like GBIF often have the benefit of giving you way more data to work with than any individual researcher could ever collect. In this type of modeling, this volume of data has a 'quality' of its own, offering the broad geographic representation needed to robustly define species-environment relationships. You may be better off using a massive amount of somewhat imperfect data than using a much smaller number of highly reliable occurrence points that don't fully capture the niche breadth of a species.

28. **Question: I had trained the rf/maxent model to the area where the presence points were found and then projected it using maxent to the entire state.**

Should we train also on the entire area of interest or stick to where the presence points were found mostly?

a. Using the area where the points were found (and some buffer) is a valid approach for using pseudoabsences. You may want to train your **RF** model using areas close to the presence locations to create pseudoabsences and then predicting on a larger area. You might even consider "donut" sampling—placing pseudo-absences at an intermediate distance from your presence points—to ensure they are not too close (risking false absences) nor too far (becoming environmentally irrelevant and inflating your accuracy metrics). When using **maxent** however, you are creating background points, and then you should sample the entire environmental envelope you are modeling.

29. **Question: Could you please share your experience on how efficient you find creating grids and working with grids for huge areas like central America or Australia?**
    a. We find the code quite efficient in creating the grid for the blocks. In the example we showed, we created a grid across the entire South America with no problem. So, I do not expect it to be a problem for Australia. The problem could be if your grid size is too small, then handling too many blocks may become more memory intensive.

30. **Question: If the species we study live in urban areas, can we add artificial structures to the model analysis if they also provide suitable habitat for them?**
    a. Yes, definitely. If you can represent an important predictor variable as a raster dataset then you can use it in your model. That is exactly when knowing the ecology of the species allows you to select adequate predictor variables. You probably want to make sure you have the adequate grid size. A pixel size of 1km2 will probably lose too much information in the urban setting but a pixel size of 30 or 10 m may be better.

31. **Question: When should we use model average and modes?**
    a. Hopefully we answered this point to some extent in the presentation. There are definitely different perspectives on when and whether to use model averaging. Generally, you would use model averaging to avoid overfitting that may occur with a single algorithm or a single training

sample. However, the downside of model averaging is that the average may rarely outperform the best individual model.

32. **Question: Slide 51 points out that many algorithms cannot extrapolate outside their training range, recommending a MESS analysis for future projections. From a biological standpoint, when a future emission scenario introduces novel climate conditions (non-analog climates) completely absent from our baseline calibration, how should we ecologically interpret the HSI (Habitat Suitability Index) values in those flagged pixels?**
    a. It is a good practice to show those areas in the results where caution is needed at interpreting the model results because the range of predictor variables is outside the range of values used to fit the model. Barbet-Masin et al. 2012 recommend using an ensembling technique to also account for variation in predictions by different algorithms. In these cases, it is also just good practice to acknowledge the fact that future suitability may not be accurately predicted based on the current distribution and climate/habitat associations of the species.

33. **Question: In slide 46, you show that moving from a localized park scale to a continental scale completely alters variable importance, inflating the perceived role of broad climate bands. For climate change mitigation, if we use regional climate predictors to build maps, don't we risk missing microrefugia and localized canopy-buffering effects that operate below the 1000m resolution?**
    a. The map itself would ideally be built using both regional and local-scale variables, so it should not completely “miss” those local effects. However, a change in variable importance is a common statistical artifact of shifting analysis scales: as your analysis shifts from local to regional scale then you are likely to see variable importance scores shift as well.

34. **Question: If I want to do SDM for a small landscape around 600 km2, then at what resolution should my environmental variables be?**
    a. The resolution of your analysis is dependent on your research question of interest and the ecology of the species you are trying to model. What you really need in a SDM is a difference between your occurrence dataset and your absence or pseudo-absence data.

35. **Question: When projecting for future climate scenarios the set of predictors become limited. Can you please suggest these predictors covering large spatial scales and resources for the same?**
    a. Google Earth Engine does provide some future climate scenarios, for instance the Terraclimate Individuals years for +2C and +4C climate futures. These can be found under the Water and Climate layers on the awesome-gee-community-catalog and ultimately in the GEE catalogue. Current and projected climate data for North American (CMPI6 scenarios) are also provided. But you are correct, there will be limitations on the availability for future scenario modeling. Note that uncertainty increases with these predictions as the time period increases from current conditions into the future.

Part 2 Question & Answer Session B

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

1. **Question: When defining spatial extent for the model, is there a way to account for what are likely to be dispersal-related absences (especially for invasive species)?**
    a. SDMs are tricky when you have dispersal-related absences since these models implicitly assume that species distributions are determined solely based on environmental predictor variables. I think dispersal-limited absence causes two issues: 1) we risk classifying unreachable but suitable habitat as unsuitable because there are no occurrence points there to inform the model and 2) we may map the distribution of a species in suitable habitat that they cannot actually reach. A simple option is to restrict your study area to just an area within some buffer distance of your occurrence points (this buffer could represent the species' maximum potential movement). You can also do a two-part approach of first modeling potential habitat and then applying a secondary filter based on distance from known occurrence points. There are other approaches too - like using more mechanistic models that actually simulate species spread dynamics - but I think the key point here is that you are dealing with a special case that needs a non-standard SDM approach.

2. **Question: Following up on the question about dispersal limitation, is there a way to remove pseudoabsences from consideration based upon some metric of similarity to presence pixels?**
    a. Hopefully this was answered in the presentation. But you can use an environmental profiling technique to create pseudoabsences in areas within your area of interest that are dissimilar to the areas where your presence data is. So, in a way you are limiting the areas to create pseudoabsences rather than eliminating them.

3. **Question: Is there any concern with machine learning models regarding multicollinearity among predictors (e.g., WorldClim biovars and DEM-derived variables)?**
    a. Some SDM models are fairly robust to collinearity (e.g., random forest) while others are not (e.g., logistic regression). However, it is often a good idea to check for highly correlated covariates as a preliminary step. It is also worth noting that variable importance metrics can be influenced by collinearity since having a bunch of correlated variables in a model tends to reduce important statistics for these variables.

       I would say it is best practice to not blindly include all WorldClim variables just because they are there, but to really have a reason to include each one (e.g., Bio 6 is important because the coldest temperatures relate to winter mortality). More recently, I have seen researchers running a first model with all predictor variables, then looking at the variable importance and running the final model with the top predictor variables.

4. **Question: How should I define the study area in Google Earth Engine when investigating recent species range expansion? For example, for an invasive species, or a species moving into areas that were previously unsuitable, should the region of interest include only the known distribution, or also surrounding areas that may be newly colonized? What criteria are recommended for identifying and selecting these potential expansion zones?**
    a. You can set up an area of interest for training your model and then predict on a different area. In this example, you can fit your model for the area the species actually inhabits and then test the potential habitat in another place (an area where the species could potentially invade). One thing to consider is making sure the area where you predict is within the variation of values used for model training.

5. **Question: Would it be possible to pick out pine tree cover with the 1m resolution canopy height layer or is there something high resolution for pine tree cover over a 30-acre study area?**
    a. The 1m canopy height and cover are not likely to distinguish species. You may be able to train a machine learning model to potentially identify species, but you will need other useful predictor variables of similar high

resolution. There are remote sensing studies where people use multi-spectral or hyper-spectral data to map unique tree species, so I would recommend doing some literature review to see what some of the top approaches might be.

6. **Question: If I am interested in a group of species, let's say benthic fishes vs. pelagic fishes, is it possible to apply this to several species at the same time to compare?**
    a. I, Ramiro Crego, modeled vegetation communities for my master thesis. If the ensemble of species responds similarly to environmental variables, you can. For instance, I modeled meadows across arid Patagonia, so a group of species with very specific niches. However, if the species have very different niches, the algorithm will struggle to find patterns. In that case it may be better to run a model for each species and then combine all those predictions. There has also been a lot of interest/research on Joint Species Distribution Models (JSDMs), which can be really powerful because information can be shared across species to improve predictions based on patterns of co-occurrence and even account for the potential for biotic interactions to influence distributions. I don't know of any way to fit these more sophisticated JSDMs in GEE or with machine learning algorithms.

7. **Question: I am interested in a spp which only lives in the flooded areas, which product could be useful for define the suitable habitat?**
    a. That is a difficult question for someone with no expertise in the species of interest. The best approach is to read the literature to see what defines the niche of the species and then try to identify predictor variables that correspond to those. People often use SAR radar data to map flooded areas and there is also a Landsat-based Global Surface Water dataset in the GEE Catalog. Percentage of sand in the soil could also be a good potential predictor variable.

8. **Question: I'm interested in spatial ecology using GEE—not only niche modeling, but also other applications like macroecology, landscape ecology, and movement ecology. How can I find more spatial ecology resources in GEE (which keywords, researchers, labs, courses, or workshops should I look for)? Also, can I expect more niche modeling algorithm implementations in GEE in the future?**

a. Well, we are a bit biased, but NASA ARSET provides several other valuable courses, including the Integration of Animal Tracking and Remote Sensing. You can browse the full list of trainings.

b. Dave Theobald has also developed a GEE toolbox for calculating various landscape metrics.

c. The Smithsonian-Mason School of Conservation also provides graduate and professional short courses.

9. **Question: Can I use GBIF data to carry out SDM research?**
    a. Yes, definitely. In fact, many studies published in the literature use citizen science data. The example we showed uses data from GBIF. It is good to be aware that studies with GBIF data typically include a number of cleaning/filtering steps. For instance, you can filter for records with low coordinate uncertainty, remove obvious outliers, apply spatial thinning to minimize the effects of sampling intensity bias, etc...

10. **Question: If I have real absences, places we went to look for the spp and we didn't find them. How can we include them in our model?**
    a. Yes, this is the best-case scenario. If you look into the original paper's code, we included code to run models with presence and absence data.

11. **Question: Is target-group method a good practice for creating pseudoabsence?**
    a. Yes, it is a very good approach if you have the data for it. For researchers working with plants, it is common to have lists of species from specific expeditions or sites, and if your species is not listed there, you can treat those locations as absences.

12. **Question: When generating training pseudoabsence points you mention that there is a command related to number of grid cells. Is that an actual value that needs to be written in, or is that extracted from a raster that was defined to serve as the area from which pseudoabsences can be drawn/created? Sorry; I'm not understanding that part of the process.**
    a. You need to specify a custom grid as well as the proportion of cells in the grid that you want to use for training the model and for testing. A

common number is 70% for training and 30% for testing. You can also use 80% for training, 20% for testing. You set that number in line 266 of the code:
// Define partition for training and testing data
var split = 0.70; // The proportion of the blocks used to select training data

13. **Question: When generating a binary image from a suitability model, what metric is used to determine the presence vs. absence threshold value? Or does the user define that?**
    a. This was hopefully responded to in the video. In this example we are using two approaches: the map shown is the result of the mode among the 5 iterations of the classification provided by the random forest. A more adequate approach is to use the threshold that maximizes the sum of specificity and specificity. Doing this is memory intensive so we only provide the code to export the resulting map in batch mode.

14. **Question: I am doing research on modelling of plant species in UK using SDM and GBIF data source. Initially I downloaded 34,616 and removed duplicated points. Applying 1 km spatial thinning, I retained 3,627; 5 km: 1468; 10 Km: 727. Which spatial thinning is suitable for my modelling?**
    a. It is very nice to see you are making progress with the code and using your own data. The question you should ask is what is the best spatial resolution for the analysis. Then use that spatial resolution for thinning the data. It's worth knowing that there is some debate in the SDM literature - with some arguing that thinning reduces spatial autocorrelation and others cautioning that it discards valuable information. We discuss this briefly later in the presentation, but a minimal level of thinning is to ensure you have no more than one presence point in each grid cell of your predictor variable stack. Thinning more than that and you are balancing information loss against the consequences of overfitting to your sample if it is spatially biased.

15. **Question: I am not sure I understand the Spatial Block Cross-Validation slide. If 70% of the cells in the study area are used to train the model, do you use all the presence and pseudo-absence points that fall within those training cells? I'm asking because, in the slide, I see multiple points within a single cell. That makes me wonder whether the split is based on cells or**

on individual points, since those would result in different numbers of training and validation samples.

a. The example we gave in this training divides your presence and pseudo-absence points into larger spatial blocks to perform a “Spatial Block Cross-validation”. So, you select 70% of the spatial blocks and use all presence and pseudo-absence points from that subset of blocks to fit the model. The idea then is to test how well your model has learned underlying ecological relationships by attempting to predict the class for any points that were withheld in the 30% of spatial blocks used for validation.

16. **Question: I have a dataset that actually includes both presences and absences, but absences truly are more common than presences (the magnitude of this difference varies by species but can be 2x-1000x). Will this be a problem when evaluating fit with AUC?**
    a. For the very unbalanced datasets, the AUC-ROC will be a problem and can give you a false impression of high accuracy. A better approach will be to use the AUC-PR. Keep in mind that modeling rare species can be tricky. You have to make sure that when you split the dataset for fitting and validation, that you have enough presences after they get split randomly. There are joint species distribution modeling approaches that treat each species as a random effect or estimate a correlation matrix among species and can improve predictions for rare species. I don’t know of any way to fit these more sophisticated JSDMs in GEE or with machine learning algorithms (you can’t do this with basic GEE code with either the Code Editor or Python API).

17. **Question: Do you know if there are some papers that have tested the approach described here (RF using training presence data and the resulting problems with accuracy assessment/validation of the model) versus a MCDA approach (such as AHP which also has potential problems with user-defined weights, but these can be assessed and variable sensitivity can be tested)?**
    a. We are not familiar with MCDA or AHP.

18. **Question: What is the general ratio of absence:pseudoabsence**
    a. In the example using Random Forest, we are generating balanced datasets with a similar number of presence and pseudoabsences for

each model iteration, following results from Barbet-Massin et al. 2012 (DOI: 10.1111/j.2041-210X.2011.00172.x) who showed that balanced datasets often give better performance in machine learning algorithms. A 1:1 ratio is a common starting point but is not a universal rule. For example, if you have a narrowly distributed species with a large study area containing a lot of unsuitable habitats, then it can be better to sample a much larger number of background points.

19. **Question: When expressing variable importance, how is 'importance' determined? For example, by holding all other variables at their mean/median and seeing influence of focal variable? Or some other method?**
    a. With the Random Forest classification in GEE, these are Gini importance values that - in a nutshell - measure how much each predictor variable contributes to sorting presence and absence points across all the decision trees that make up the Random Forest. You could engineer a workflow for assessing variable importance by permutation and comparing performance metrics with/without a given predictor variable.

20. **Question: Can collinearity (or an equivalent) of predictors occur with random forest models? If so, how can we detect this?**
    a. Machine learning algorithms are fairly robust to collinearity (e.g., random forest). Collinearity is a problem when using parametric approaches (e.g., logistic regression) and the goal is to make inference. That said, you can check for highly correlated covariates as a preliminary step. It is also worth noting that variable importance metrics can be influenced by collinearity since having a bunch of correlated variables in a model tends to reduce importance metrics for these variables.

       In the original code we published, there is code to check the correlation among predictor variables.

21. **Question: How can we remove heavily clustered presence data?**
    a. Typically, you would use some level of spatial thinning, which we implemented in our example by defining a grid and keeping one point per grid cell. See also Question 14 which includes some discussion of thinning strategy.

22. **Question: What should I do if I have about a hundred true absences, but I have several thousand presence points? Do I still use the absences?**
    a. In that case you will have a very unbalanced dataset. One approach could be to randomly select the same amount of presence than absences. You may have enough data to fit good models and use the extra presence data to validate your model predictions. You can run several iterations using a different random set of presence data and then average the outputs. You can also fit a model with pseudoabsences using all your presence data and then compare model outputs.

23. **Question: If we take these codes and adapt them in our work, should we cite NASA or another source?**
    a. The SDM workflow presented here is derived from one of the case studies we published in a paper, so that would be the main reference.