Implementation of species distribution models in Google Earth Engine

Ramiro D. Crego1,2 | Jared A. Stabach1 | Grant Connette1,2

Abstract
Aim: Google Earth Engine (GEE) is a free Web-based spatial analysis platform that requires only a web browser and an Internet connection to programmatically access and analyse data from its multi-petabyte catalog of regularly updated satellite imagery (e.g. MODIS, Landsat, Sentinel) and other geospatial datasets. The high computing capacity of GEE can make computationally demanding analyses more accessible to researchers and practitioners, especially those with limited access to advanced computational resources. Here, we present a workflow in GEE to fit species distribution models, offering direct access to a multi-petabyte catalog of raster products to obtain estimates of habitat suitability.

Innovation: We implemented a workflow for species distribution modelling in GEE that includes importing species occurrence data into the GEE platform, selecting and preparing predictor variables, and performing model fitting with spatial or temporal split-block cross-validation techniques. We present three case studies that demonstrate: (i) a baseline SDM workflow that produces informative model predictions, (ii) a workflow that accounts for temporal variability in predictor variables to study changes in habitat suitability over time and (iii) a complex and computationally demanding analysis incorporating thousands of satellite images for modelling habitat suitability at high spatial resolution.

Main Conclusions: Our SDM workflow allows users to benefit from the high speed and performance of GEE without the need for significant computing infrastructure. This workflow may be especially beneficial to researchers in countries where computing power is limited, as SDMs frequently require the download, storage and processing of large raster datasets. We also discuss key limitations of implementing SDMs in GEE, such as user memory limits and the lack of high-level functions. We include a step-by-step guide for the general model workflow and for each of the case studies presented to facilitate its implementation.

Keywords
gradient boosting, habitat suitability models, JavaScript, niche models, occurrence data, random forest
1 | INTRODUCTION

Understanding where species occur in space and time is a fundamental question in ecology. This basic question has motivated the development of species distribution models (SDMs), also known as ecological niche models or habitat suitability models, that link known species localities with predictor variables to assess patterns of species occurrence and habitat suitability (Guisan & Thuiller, 2005; Guisan & Zimmermann, 2000). These models have become a critical component of studies in the fields of ecology, evolutionary biology, wildlife management and conservation (Guisan et al., 2017; Pecchi et al., 2019; Zimmermann et al., 2010). Currently, SDMs are the most widely used modelling tool for analysing occurrence data and predicting habitat suitability (Araújo et al., 2019).

The increase in the use of SDMs is related to the advance in computing power and software to implement a large variety of algorithms (Guisan et al., 2017; Kass et al., 2018; Phillips et al., 2006, 2017; Thuiller et al., 2009). Currently, algorithms such as generalized linear models, generalized additive models, random forest, gradient boosting and maximum entropy (among others) are commonly used in SDMs (see Guisan et al. (2017) for a complete guide to SDMs). The increased variety of statistical models available to researchers has also been matched by tremendous advances in remote sensing as a scientific discipline, making available a vast array of freely accessible datasets that can be used as predictor variables (Oeser et al., 2020; Pettorelli et al., 2014; Schulte & Pettorelli, 2018; Xiong et al., 2017). However, accessing and manipulating large raster datasets, many of which are organized across different sites/sources, require significant computation power and processing time, which may be limiting factors for many researchers, conservation practitioners and/or land managers.

Google Earth Engine (GEE) was launched as a platform that could process and analyse a wide array of satellite imagery and geospatial datasets, making computationally demanding workflows more accessible. This cloud-based spatial analysis platform is hosted by Google and is free to all users for research purposes, providing access to high-performance and an intrinsically parallel computing system that subdivides computations across the Google computational infrastructure, speeding up computational processes (Gorelick et al., 2017). Google Earth Engine has created new research possibilities for remote sensing scientists, with thousands of datasets and algorithms in a single online platform, reduced data processing timelines and improved computational efficiency (Tamiminia et al., 2020). Many of the algorithms used by remote sensing experts are the same algorithms used in SDM modelling and other spatial ecology analyses. Therefore, GEE can make computationally demanding analyses more accessible to researchers and practitioners interested in spatial ecology, especially those located in countries where computational resources are more limited.

Here, we present a species distribution modelling workflow implemented in Google Earth Engine. We first explain the basic workflow and then demonstrate its implementation using Bradypus variegatus as a case study, a species widely used to present other SDM software and R packages (Hijmans et al., 2017; Kindt, 2018; Phillips et al., 2006, 2017). We then showcase several advantages of implementing SDMs in GEE with two additional case studies. In a second case study, we account for temporal variability in predictor variables through an example that examines change over time in suitable habitat for Cebus capucinus. In a third case study, we use Landsat 8 Surface Reflectance (SR) and Advanced Land Observing Satellite Synthetic Aperture Radar (SAR) mosaics as the basis for mapping breeding grounds for Hylocichla mustelina across the eastern continental USA at 90-m spatial resolution. We also publish a full tutorial with this manuscript that covers a step-by-step modelling workflow for SDM analyses in GEE, from how to import occurrence data to visualizing and exporting results.

2 | MODELLING WORKFLOW

Species distribution models have two prominent applications, characterizing the predictors that define a species’ niche and projecting fitted models across a landscape or seascape to map and visualize suitable habitats (Araújo et al., 2019). The modelling workflow we present here employs non-parametric algorithms available in GEE for mapping suitable habitats for single species. The steps required to obtain final model predictions include importing occurrence data into the GEE platform, defining the spatial resolution and extent of the analysis, selecting and preparing predictor variables, and performing model fitting with spatial or temporal split-block cross-validation techniques (Figure 1). The general modelling workflows and all the code used for the case studies are described in detail in Appendix S1, with associated public repositories that include code that can be adapted to new datasets. While code can be completely customized for different needs, we developed a general workflow that will work with any new dataset while requiring minimal edits for users with little JavaScript programming experience.

2.1 | The response variable

Obtaining good quality data is central to the modelling process (Araújo et al., 2019; Leroy et al., 2018; Sillero et al., 2021). Once species location data have been acquired, it must be imported into GEE as an “asset” from a csv file or shapefile (Appendix S1 as Figure S1). Datasets that include both presence and absence data are always recommended for use in SDM analyses (Brotons et al., 2004). However, often only presence data are available, especially when location data are accessed from museum records, citizen science data or online databases such as the Global Biodiversity Information Facility (GBIF) or iNaturalist. In these cases, random background or pseudo-absence points must be generated to fit a model (Phillips et al., 2009). Choosing the proper method to select pseudo-absences is critical as it can affect model performance (Barbet-Massin et al., 2012). In our general workflow, we provide code to implement three different methodologies to generate
pseudo-absences, including: a) random pseudo-absences across the entire study area; b) random pseudo-absences within a specified distance from presence locations—limiting pseudo-absences to areas potentially accessible to the species (Araújo et al., 2019) and to account for the potential geographical or environmental sampling bias (Phillips et al., 2009); and c) implementing an environmental profiling technique to limit the area to select pseudo-absences by masking out known environmentally suitable locations derived from a clustering method, similar to other two-step methods (Chefaoui & Lobo, 2008; Senay et al., 2013). The geographical or environmental stratified selection of random pseudo-absences should be preferred to use with machine-learning algorithms available in GEE (Sillero et al., 2021).

Most of the datasets obtained from global databases contain important sampling biases that can hamper model predictions (Guisan et al., 2017; Sillero et al., 2021). We implemented a spatial thinning approach to reduce the potential for clustered presence locations to result in models with sampling bias, which has been shown to result in species occurrence data that provide better model performance (Boria et al., 2014; Fourcade et al., 2014; Veloz, 2009). The location data are thinned to one randomly selected occurrence record per grain size at the chosen spatial resolution.

### 2.2 Spatial extent, spatial resolution and predictor variables

The area of interest can be drawn automatically as the extent that encloses all species locations, can be defined by a polygon or can be drawn directly on GEE. Ecological patterns are scale-dependent, and choosing an appropriate study area extent and spatial resolution of the analysis is critical (Chave, 2013). The extent of the analysis should be carefully selected and constrained to a realistic realm of the species of study, avoiding unrealistic extents that can hamper
model accuracy and predictions (Guisan et al., 2017; Leroy et al., 2018; Sillero et al., 2021).

Frequently, studies with large spatial extents use coarser spatial resolution focused on climatic predictors, whereas studies with more local extents use finer spatial resolution often focused on habitat predictors (Pearson & Dawson, 2003). However, both types of variables can be important across scales (Austin & Van Niel, 2011). A key limitation for performing fine-resolution analysis at large scales is the increased demand for computing power. GEE was developed to handle large-scale (e.g. planetary) raster operations and makes it possible to perform many analyses at large extents with high spatial resolution.

Google Earth Engine also provides direct access to the curated multi-petabyte data catalog, which increases at a rate of 6000 new scenes per day from active satellite missions (Gorelick et al., 2017). Refined predictor variables (e.g. forest cover, population density, night lights and human modification index) and long-term image collections (e.g. Landsat, MODIS and Sentinel) are all available. These data layers can be used in their basic format or analysed to create new data layers (e.g. land cover classifications), all of which can be incorporated as independent variables in modelling frameworks to predict species occurrence.

### 2.3 Classification algorithms and model validation

Several non-parametric classification algorithms are available in GEE that have seen widespread use in habitat suitability modelling. These include random forest, support vector machine, classification and regression trees, maximum entropy and gradient boosting.

Ideally, model outputs should be validated using truly independent data (Araújo et al., 2019). Often, however, such data are not available. As a result, it is common practice to partition available location data into training and testing sets. In our main workflow (see Case Studies 1 and 3), we implemented a spatial block cross-validation technique to partition data for model training and model validation (Roberts et al., 2017; Valavi et al., 2019). This spatial blocking technique is preferred to random partitioning of point locations for cross-validation as the latter tends to inflate accuracy metrics (Roberts et al., 2017). Users can run multiple iterations with random block splits to create different training and validation datasets. Pseudo-absences are randomly created within each set of training and validation blocks at each iteration (Appendix S1 as Figure S8). We also implemented a temporal block cross-validation technique in which certain years of data are used for model training and a separate year of data is used for model validation (see Case Study 2). In both spatial and temporal block cross-validation approaches, the same number of pseudo-absences as presence data is created at each iteration, resulting in a balanced dataset that has been argued to perform better with machine-learning techniques (Barbet-Massin et al., 2012; Elith et al., 2008; Evans et al., 2011; Sillero et al., 2021).

In our workflow, model accuracy based on the validation partition for each iteration of model fitting is estimated. The first metrics are the threshold-independent area under the curve of the receiver operating characteristic curve (AUC-ROC; Fielding & Bell, 1997) and/or the area under the precision-recall curve (AUC-PR; Sofaer et al., 2019). The AUC-ROC ranges from 0 to 1, with 1 indicating perfect discrimination between presence and absence–pseudo-absence locations, 0.5 discrimination is no better than random performance and <0.5 performance is worse than random. The AUC-PR also ranges from 0 to 1, with 1 indicating perfect prediction of presence data, while random performance is equal to the prevalence of presence locations in the dataset (Sofaer et al., 2019).

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 increases 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).

The second metrics reported are based on the confusion matrix: sensitivity (correct predictions of the occurrence) and specificity (correct predictions of the absence; Fielding & Bell, 1997). These metrics require defining a threshold. For each iteration, we used the threshold that maximizes the sum of sensitivity and specificity. This threshold has been shown to perform well with presence-only data (Liu et al., 2016). Note that all information of the confusion matrix can be obtained if the analyst wishes to calculate other threshold-dependent metrics, such as the true statistic skill or Sørensen’s similarity index (Leroy et al., 2018).

### 2.4 Model predictions

Final habitat suitability maps for a given classifier are calculated as the average presence probability across all iterations of model fitting with different sets of pseudo-absence data. There are two different methods for obtaining final binary presence–absence predictions in our workflow, which are related to the way in which GEE operates. Google Earth Engine can compute operations in two modes, on-the-fly and batch (Gorelick et al., 2017). The on-the-fly mode runs computations interactively and displays results immediately in the console or on the map. This mode, however, has a 5-minute time-out limit, after which a memory limit error is displayed (Navarro, 2017). Contrastingly, the batch mode can run computations indefinitely, as long as the process continues to make progress (Navarro, 2017). This allows the user to run longer operations that will fail to execute in the on-the-fly mode (Navarro, 2017). Consequently, binary maps...
can be calculated using on-the-fly computations to rapidly display the resulting map or using the batch mode for a more computationally demanding method. In our workflow, the on-the-fly binary map consists of a majority vote for the classifier output (presence or absence) among all the selected iterations from the chosen classifier. This method is more efficient, and the results are rapidly displayed in the GEE map viewer. The batch mode consists in generating a binary map from the habitat suitability model using the mean of the thresholds that maximized the sum of sensitivity and specificity for each of the individual model predictions. This operation is more computationally demanding, and depending on the number of presence points, the spatial resolution chosen and the number of model iterations, it may reach the 5-minute time-out limit for an on-the-fly mode operation in GEE. To obtain the binary output map using this threshold approach, the user may use the batch mode on GEE to export the output to Google Drive. Similarly, all desired model outputs can be exported by the user, including accuracy assessment metrics for complex models (see Case Study 3).

3 | CASE STUDIES

3.1 | Modelling species distribution using presence-only data

We obtained occurrence data for *Bradypus variegatus* from GBIF (GBIF.org [20 January 2021] https://doi.org/10.15468/dl.jxcv7e). We filtered data to obtain georeferenced records between the years 2000 and 2020 that had a coordinate uncertainty <250 m. We further cleaned the dataset by manually removing all locations that fell on top of buildings or water bodies, assuming incorrect coordinates. We set the spatial resolution of the analysis to 1 km. After randomly removing all but one location within each 1-km pixel, our presence-only dataset included 323 occurrences of the species. We used a combination of climatic predictor variables (temperature seasonality, maximum temperature of warmest month, minimum temperature of coldest month and annual precipitation; Hijmans et al. (2005)), elevation (Shuttle Radar Topography Mission (SRTM); Farr et al., 2007) and percentage tree cover obtained from the Terra MODIS Vegetation Continuous Fields (VCF). The VCF product is generated yearly and produced using monthly composites of Terra MODIS Land Surface Reflectance data. We calculated mean percentage tree cover for the period corresponding to the actual dates of the occurrence dataset (2003–2020).

We implemented a repeated (10 times) spatial block cross-validation technique, randomly partitioning 200 x 200 km spatial blocks for model training (70%) and validation (30%) while randomly generating pseudo-absences at each iteration. Each training dataset was fit with a random forest classifier with 500 trees (Evans et al., 2011). To maintain a balanced dataset, which is known to perform better with certain non-parametric classification techniques such as random forest (Barbet-Massin et al., 2012; Elith et al., 2008; Evans et al., 2011; Sillero et al., 2021), we generated an equal number of pseudo-absences as occurrence data. We used a two-step environmental profiling approach to restrict the area for the creation of pseudo-absences (Chefaoui & Lobo, 2008; Senay et al., 2013). We first performed a k-means clustering using Euclidean distance to divide pixels in the study area into two clusters: one cluster with similar environmental characteristics to a random subset of 200 presence locations and a second cluster with dissimilar characteristics. We then created random pseudo-absences within the boundaries of the cluster more dissimilar to the presence data. We assessed model accuracy by calculating AUC-PR, AUC-ROC, sensitivity and specificity on the validation dataset.
dataset of the 10 iterations. We generated a final habitat suitability map by averaging the model-based presence probability outputs across all iterations. We generated a final predicted presence–absence map by using a majority vote for the 10 binary model outputs. See Appendix S1 for a step-by-step description of the code.

Final predicted suitable habitat (Figure 2) was qualitatively similar to those reported by various authors for the same species (Hijmans et al., 2017; Kindt, 2018; Phillips et al., 2006, 2017). The mean AUC-PR (mean: 0.83, standard deviation (SD): 0.05) and the mean AUC-ROC (mean: 0.91, SD: 0.06) from the 10 model iterations indicated good model performance and were in the range of the AUC-ROC reported by previous research using a similar dataset implemented in Maxent (Phillips et al., 2006) and the BiodiversityR package (Kindt, 2018). Mean sensitivity was 0.90 (0.07 SD), and mean specificity was 0.81 (0.12 SD) for the 10 model iterations.

3.2 | Accounting for temporal dynamics in forest cover

One main limitation of most SDMs is that temporal resolution is not considered when modelling habitat suitability and species distributions (Araújo et al., 2019). This is particularly important when using species location data acquired across several decades in temporally dynamic landscapes. Ideally, the acquisition date of predictor variables should match the date at which species location data were collected to provide information across seasons or years, rather than using annual or multi-annual averages (Araújo et al., 2019). We used Cebus capucinus as an example to demonstrate a modelling framework that takes advantage of GEE to match each occurrence record to predictor variables extracted from the satellite image with the nearest acquisition date in a time series. The high preference of the species for primary and secondary forests (Mittermeier et al., 2013) in areas suffering high rates of deforestation (Hansen et al., 2013) makes it a good example to illustrate the use of long-term satellite data to model habitat suitability changes across time.

Similar to the steps described above, we obtained occurrence data for Cebus capucinus from GBIF (GBIF.org (27 January 2021) https://doi.org/10.15468/dl.qus4ha) and retained only georeferenced records with a coordinate uncertainty <250 m within Panama and Costa Rica (note that some authors divide this species into Cebus capucinus and Cebus imitator; Mittermeier et al., 2013). We used mean annual temperature, annual precipitation (Hijmans et al., 2005), elevation (Farr et al., 2007) and percentage tree cover (Terra MODIS VCF, 250 m resolution) as predictor variables. We set the spatial resolution of the analysis to 250 m to match the MODIS data. We used data from between 2003 and 2018 for model fitting and thinned the data to no more than one occurrence record per year from each 250 m pixel. For each species location record, we then identified the Terra MODIS VCF image closest in time to that observation and extracted the percentage tree cover value. Our final dataset contained a total of 330 occurrence records.

To create a matching pseudo-absence location for each occurrence location, we generated a random pseudo-absence point within a 100 km buffer, extracting elevation, mean annual temperature, annual precipitation and the per cent tree cover from the VCF image that corresponded to the time period of the occurrence point. We repeated this process five times, resulting in five balanced datasets for each iteration, each with a different set of pseudo-absences. We fit a random forest model (500 trees) to each individual dataset and then averaged model outputs to calculate the mean habitat suitability of the five iterations.

Because model predictions will vary each year due to changes in underlying predictor variables, we withheld data from 2019 for model validation. We selected the year 2019 for our out-of-sample model validation because this year had a large number of occurrence records and was the last year the MODIS VCF product was available in GEE. We assumed that if the model predicts well for data in 2019, then the model likely performed well in other years and is useful for predictions. We used 125 occurrence records, and a set of 125 pseudo-absences randomly created across the study area (restricted to a 100 km buffer around each occurrence location) to estimate the AUC-PR, sensitivity and specificity for each of the five individual model predictions using the percentage of tree cover for 2019, together with mean annual temperature, annual precipitation and elevation as the predictor variables.

After validating model predictions for 2019, we used the five trained models to predict habitat suitability across the entire MODIS VCF image collection (2000–2019). We then applied the GEE `Trend` function to the image collection of predicted annual habitat suitability. This function fits a linear regression on a pixel-by-pixel basis to the 20-year time series of predicted habitat suitability values. The output of this function is an image (raster) with two bands indicating the slope of the linear regression and a t test statistic for the significance of the trend in the time series at each pixel. This output can be used to identify areas where habitat suitability increased or decreased across the 20-year study. See Appendix S1 for a step-by-step description of the GEE code.

The models accurately predicted the withheld 2019 data, with a mean AUC-PR of 0.81 (SD: 0.06), mean sensitivity of 0.78 (SD: 0.03) and specificity of 0.84 (SD: 0.05) across the five-model iterations fit with different pseudo-absence points. Our results predicted Cebus capucinus habitat suitability across 20 years and identified trends in habitat suitability for the last two decades (Figure 3). The reduction in suitable habitat closely corresponded to deforested areas identified using Landsat imagery between 2000 and 2019 (Hansen et al., 2013; Figure 3). These results show the flexibility of the modelling workflow to account for temporal dynamics in SDMs when data are acquired across a long time period by obtaining time-matched values of a temporally varying predictor variable. This modelling framework can be extended to use MODIS 16-day NDVI and EVI product datasets, which may be particularly relevant for modelling habitat suitability of migratory species for which habitat conditions change seasonally. Resulting maps could be useful to inform conservation programmes by identifying seasonal changes in habitat suitability.
Modelling species distribution at high spatial resolution using unclassified satellite images as predictor variables

Recent studies have demonstrated the benefits of using unclassified multispectral satellite images as predictor variables to model the habitat suitability of species (Shirley et al., 2013; St-Louis et al., 2014). Using unclassified multispectral data to model species distributions has the benefit of avoiding the errors of misclassified pixels in land cover products and the amount of information lost when classifying continuous data into discrete subjective categories (Bradley & Fleishman, 2008). However, processing large numbers of multi-band satellite images is a highly demanding computing process and working with multiple satellite images often requires several gigabytes of data download and storage. To demonstrate the implementation of the SDM workflow in GEE combining presence records and unclassified satellite data, we modeled the breeding distribution of *Hylocichla mustelina* across the eastern continental USA. *Hylocichla mustelina* is a common bird that prefers to breed in deciduous and mixed forests, within patches as small as 1 ha (Hoover et al., 1995).

We obtained the eBird observation dataset for *Hylocichla mustelina* from GBIF (GBIF.org) for the month of June (middle of breeding season when observation of migrants is more unlikely) for the years 2018, 2019 and 2020. We set the spatial resolution of the analysis to 90 m. We rarefied the original 99,939 observations to keep no more than one observation per pixel, resulting in 34,880 observations for modelling. We defined the extent of

![Figure 3: Changes in habitat suitability in the last two decades for *Cebus capucinus* (2000–2019). The main map shows the regression slope calculated from 20 years of habitat suitability values. Red pixels are areas with declining habitat suitability, and blue pixels are areas where habitat suitability increased across years. Only pixels with significant trends are shown (α = 0.05). This output was obtained by fitting a linear regression to habitat suitability values for each pixel across 20 years of data. The zoom-in boxes show from bottom left to bottom right: predicted habitat suitability for 2000, predicted habitat suitability for 2019, regression slope (i.e. direction of change in habitat suitability) and the deforested areas (in red) obtained from Hansen et al. (2013).](image-url)
the analysis to the eastern continental USA (4,606,284 km²), limiting the extent to the western observation in the dataset (longitude: -103 degrees). We modelled Hylocichla mustelina habitat suitability using atmospherically corrected Landsat 8 surface reflectance (SR) collection 2, Synthetic Aperture Radar (SAR) and mean temperature products as predictor variables. We filtered the Landsat 8 SR time series to keep only images between the months of April and August for 2018, 2019, and 2020. We included those months to obtain a cloud-free mosaic for the entire study area. For each of the 6,150 resulting Landsat 8 images, we masked out bad-quality pixels with clouds, shadow clouds and saturated pixels and rescaled pixel values with the appropriate scaling factors (https://www.usgs.gov/core-science-systems/nli/landsat/landsat-collection-2-level-2-science-products) using a predefined mask function available in GEE. We selected the blue, red, green, near-infrared and shortwave infrared 1 bands for analysis (30-m spatial resolution). For each image, we also calculated NDVI using the normalizedDifference function available in GEE. We finally created a mosaic for each band spanning the entire study area by calculating the median pixel value across all Landsat images in our filtered time series. Additionally, we used the 2018, 2019 and 2020 global mosaics of Advanced Land Observing Satellite (ALOS) Phased Arrayed L-band Synthetic Aperture Radar (SAR) HH and HV polarization datasets (25-m spatial resolution), as they are useful in identifying woody vegetation (Shimada et al., 2014). For each pixel of the two bands, we obtained the median value among the three years. Finally, to account for the variation in temperature across the breeding range, we included the mean temperature for the month of June (1-km spatial resolution; Hijmans et al., 2005). From the stack of predictor variables, we masked out all pixels containing permanent water using the global water surface product (Pekel et al., 2016).

We implemented a repeated (5 times) spatial block cross-validation technique, randomly partitioning 200 x 200 km spatial blocks for model training (70%) and validation (30%) while randomly generating pseudo-absences at each iteration. We again used a k-means clustering using Euclidean distance, based on 1,000 randomly selected occurrences, to restrict the area for the creation of pseudo-absences to pixels more dissimilar to the environmental profile of the occurrence data (Chefaoui & Lobo, 2008; Senay et al., 2013). We generated an equal number of pseudo-absences as occurrence data for each of the 5 datasets (Barbet-Massin et al., 2012; Elith et al., 2008; Evans et al., 2011; Sillero et al., 2021). We fit a random forest model (500 trees) to each individual training dataset and then averaged model outputs to calculate the mean habitat suitability of the five iterations. The total number of presence and pseudo-absence points in our five training sets varied between 40,300 and 49,260, depending on the random spatial blocks included, while each model prediction was performed across 705,147,693 pixels.

We assessed model accuracy by calculating the AUC-PR, sensitivity and specificity on the validation dataset of each model iteration. To obtain a binary potential distribution map we used the mean threshold that maximized the sum of sensitivity and specificity of each of the individual model predictions (Liu et al., 2016). We ran the model in batch mode, exporting the results to Google Drive. See Appendix S1 for a step-by-step description of the code.

Our results show predicted Hylocichla mustelina habitat suitability across the eastern continental USA at 90-m spatial resolution (Figure 4). The mean AUC-PR for the five-model iterations was 0.68 (SD: 0.01), mean sensitivity was 0.87 (SD: 0.03), and mean specificity was 0.90 (SD: 0.01). While the lack of high accuracy for the resulting map may be due to the natural bias of the occurrence data used (coordinates represent the location of the observer rather than the birds) and biases that can be amplified at such high spatial resolution, the results show the potential to model species distributions at high spatial resolution in GEE by combining thousands of occurrence data and processing thousands of images from satellite missions.

4 | DISCUSSION

Recent increases in freely available spatial datasets have been a major driver of the continued development and increased use of species distribution models (Guisan et al., 2017; Kass et al., 2018; Thuiller et al., 2009). Complex spatial analyses, however, can reach limitations of computing power, data storage capacity and long data download times, particularly in countries with less economic power. Implementing SDMs in GEE facilitates a direct link between species occurrence data, algorithms, multiple petabytes of remotely sensed data and high computing power to work with those spatial datasets. This simplifies the modelling workflow by avoiding the need for increased data storage and sufficient Internet speed/bandwidth for downloading and preparing multiple predictor variables for modelling. Our study expands on another project that uses GEE to power the display of species locations and ranges (Jetz et al., 2012). We present a workflow that follows many SDM standard practices to predict suitable habitat for species and provide model validation metrics (Araújo et al., 2019; Sillero et al., 2021; Zurell et al., 2020). We expect this modelling workflow to be most useful for practitioners who want to quickly obtain model results or for users interested in running models accounting for environmental variation across time and/or using satellite imagery to predict species distributions at high spatial resolution.

4.1 | Main advantages

Two of the biggest advantages of implementing SDMs directly in GEE are the direct access provided to the multi-petabyte data catalog and the computing resources available to the user. Users can rapidly obtain results from basic SDM analyses or can perform more complex analyses that leverage GEE’s strengths in raster processing. Running SDMs in GEE also allows users, as we demonstrate with the second case study, to incorporate the temporal variability of many predictor datasets, affording opportunities to estimate long-term trends in habitat suitability across time. It is also possible to combine
the SDM workflow with thousands of occurrence points and process thousands of unclassified satellite images to make extensive predictions of habitat suitability, as we demonstrate in our third case study. In addition to the variables used in our case studies (e.g. bioclimatic variables (Hijmans et al., 2005), elevation (Farr et al., 2007), Landsat 8 surface reflectance, surface water (Pekel et al., 2016) and ALOS-SAR (Shimada et al., 2014)), multiple other remote sensing products, such as climatic information, multiple land cover classifications and several vegetation indexes (e.g. Enhanced Vegetation Index and Normalized Difference Vegetation Index) are available in GEE. Additionally, raw satellite imagery from Landsat and Sentinel can be preprocessed to create customized land cover maps or other products, such as greenness calculated from Landsat imagery to model mammal habitat suitability (Oeser et al., 2020).

Another main benefit of the GEE workflow is the ease of reproducibility of results. Lack of reproducibility in methods and results has been cause for criticism of many published studies implementing SDMs (Zurell et al., 2020). Because all data and algorithms are compiled in an online repository, the shareable GEE script (with linked data) can be re-run by a new analyst to ensure transparency and reproducibility of results. The publicly available code can also help other researchers to spur new research.

In addition to academics, our workflow is intended to make SDM methods in GEE accessible to a broader audience. Across the globe, government agencies and conservation organizations struggle to produce reliable scientific information at the speed required to inform conservation decision-making processes (Cook et al., 2013; Cvitanovic et al., 2016). We expect that the workflow we present here will help practitioners rapidly obtain high-quality habitat suitability maps or habitat suitability change analyses across different ecosystems and taxa, helping produce the needed information for action by governments, land managers and conservation practitioners.

4.2 Workflow limitations

While GEE allows users to rapidly analyse large spatial datasets, Google imposes user quotas to prevent high resource-intensive analyses from negatively impacting the computing capacity of the
entire community (Gorelick et al., 2017). When the user exceeds the quota, an error is shown in the console of Earth Engine and the computation process stops. Thus, certain analyses that require more memory, such as fitting a presence-absence model using a defined threshold for several repeated spatial block cross-validations or running the models with large numbers of presence points (in our experience, more than 1,000 presence locations trigger memory limits), are often required to be run in batch mode to avoid reaching the computation time-out limitation of the on-the-fly mode (Gorelick et al., 2017; Navarro, 2017). This limitation restricts the user's ability to quickly display analysis results in GEE’s interactive map interface. However, these limitations can be avoided by exporting results to Google Drive or Google Cloud Storage to be displayed in other third-party software, although the process is more time-consuming (e.g., Figure 4). The amount of time required to export model outputs varies greatly depending on Internet speed, extent of the map and spatial resolution. The habitat suitability model for Case Study 3 took 22 h to run, and the validation metrics between 30 min and 11 h. The need to export results from analyses leads to another potentially important limitation of GEE, storage capacity. While Google does offer 15Gb of free storage capacity, users must be careful with how they use this free storage, as it is shared across all Google apps (e.g. Gmail, Drive, GEE), to avoid filling the free storage capacity and having to purchase additional space (Navarro, 2017). However, the final habitat suitability map of Case Study 1—a near continental-scale analysis for Bradypus variegatus—is only 52.4 Mb when exported at 1-km spatial resolution (tif format). More caution is needed with large-scale analysis at high spatial resolutions where several gigabytes of space in Google Cloud may be needed to export the results (the final habitat suitability raster for Case Study 3 was 2.7 Gb). For these reasons, we see GEE being most useful for practitioners when running models with less than 1000 presence locations, as results are obtained in matter of minutes even at continental scales. However, exporting results allows users to run more computationally demanding analyses with just a web browser and access to Internet. For this latter case, we recommend downloading the results from the cloud and deleting the large files to quickly free up the storage space.

Another limitation we encountered while implementing SDMs in GEE is the lack of efficiency in processing large vector files, such as feature collections. As a result, models that require a large number of pseudo-absence points would be more difficult to implement on-the-fly in GEE as memory limits would be quickly reached, compromising usability of the tool. This would limit a user’s ability to implement certain methods, such as inhomogeneous Poisson’s point process models (Renner et al., 2015), even if new algorithms were added to GEE in future. The balanced presence–pseudo-absence datasets implemented in our modelling framework avoid problems with memory limits, but require a moderately large presence-only sample size for analysis (Wisz et al., 2008), making our workflow more difficult to use with rare species. However, the code can be modified to increase the number of pseudo-absences (see details on Appendix S1), and if memory limits are reached, then models can be run in batch mode. Future advances in GEE may allow users to incorporate more computationally demanding techniques to create pseudo-absences.

It is also important to note that GEE does not have open-source code, in contrast with the R programming language (R Development Core Team, 2016), the most common programming language used in the field of SDMs. Working in GEE requires that users work in either the JavaScript or Python coding languages, though a recent R package, rgee, allows to bridge GEE with R software (Aybar et al., 2020). The platform also lacks many specialized functions required for spatial ecology analyses that must frequently be developed as user-created functions. This represents an initial learning curve for many researchers and practitioners, and it was the primary motivation for our presentation of a step-by-step SDM workflow and custom functions in this study.

Finally, the quality of the models will always depend on the quality of the observation data (Araújo et al., 2019; Leroy et al., 2018; Sillero et al., 2021). The main goal of our study was to present a novel workflow in GEE, and for that, we used easily accessible occurrence datasets from GBIF. However, no increase in computing power or accessibility of high spatiotemporal resolution predictor variables can remedy poor predictive performance due to biased occurrence datasets. Therefore, improving field data collection for implementing SDMs remains a high priority.

4.3 Moving forward

Google Earth Engine is a developing project, and more datasets and new algorithms are being constantly added. The Maxent algorithm (Phillips et al., 2006), for example, has been recently integrated into GEE. We expect the workflow we present here to provide a useful entry point for ecologists interested in using GEE. We also expect this platform to be applied to other complex spatial analysis problems in future, such as connectivity modelling. Finally, we hope this work will inspire Google developers to build more flexibility into many of the available algorithms to be able to manually control for model complexity. For instance, it would be useful to implement downsampling in random forest classifiers when using presence-only data (Valavi et al., 2021), but for which subsampling at the level of the individual tree is necessary.

Our workflow is intended to make SDMs accessible to a broader audience, allowing users to obtain habitat suitability maps for different species and at different scales and spatial resolutions with minimal code editing. The detailed tutorials explain the code in a step-by-step fashion. We hope that this workflow will be a valuable resource for a wide range of users, such as researchers, environmental consultants and government agencies, who are interested in quickly obtaining high-quality and easily reproducible habitat suitability and species distribution maps.

ACKNOWLEDGEMENTS

We would like to acknowledge the support of the Smithsonian Conservation Commons Working Land & Seascapes Amplification.
and Innovation Award (Award #: WLS-2020-11). We want to thank all the Smithsonian Conservation Biology Institute interns that have tested our code workflow. We thank Katie LaJeunesse Connette for providing the drawing of the Wood Thrush. We appreciate the insightful comments from the associate editor Boris Leroy, Blas M. Benito and three other anonymous reviewers, which greatly improved our work.

CONFLICT OF INTEREST
The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PEER REVIEW
The peer review history for this article is available at https://publo ns.com/publon/10.1111/ddi.13491.

DATA AVAILABILITY STATEMENT
The Google Earth Engine code and data used in this study are freely available at the following Google Earth Engine repository: https:// code.earthengine.google.com/?accept_repo=users/ramirocrego84/SDM_Manuscript
The tutorials to fit Species Distribution Models in Google Earth Engine are available in the appendix or at the following link: https:// smithsonian.github.io/SDMInGEE/

ORCID
Ramiro D. Crego https://orcid.org/0000-0001-8583-5936
Jared A. Stabach https://orcid.org/0000-0002-7896-1237
Grant Connette https://orcid.org/0000-0002-9424-6591

REFERENCES
Araújo, M. B., Anderson, R. P., Barbosa, A. M., Beale, C. M., Dormann, C. F., Early, R., Garcia, R. A., Guisan, A., Maioran, L., Naimi, B., O’Hara, R. B., Zimmermann, N. E., & Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. Science Advances, 5(1), 1-12. https://doi.org/10.1126/sciadv.aat4858
Austin, M. P., & Van Niel, K. P. (2011). Improving species distribution models for climate change studies: Variable selection and scale. Journal of Biogeography, 38(1), 1-8. https://doi.org/10.1111/j.1365-2699.2010.02416.x
Aybar, C., Wu, Q., Bautista, L., Yali, R., & Barja, A. (2020). rgee: An R package for interacting with Google Earth Engine. Journal of Open Source Software, 5(51), 2272. https://doi.org/10.21105/joss.02272
Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? Methods in Ecology and Evolution, 3(2), 327-338. https://doi.org/10.1111/j.2041-210X.2011.00172.x
Boria, R. A., Olson, L. E., Goodman, S. M., & Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. Ecological Modelling, 275, 73-77. https://doi.org/10.1016/j.ecolmodel.2013.12.012
Bradley, B. A., & Fleishman, E. (2008). Can remote sensing of land cover improve species distribution modelling? Journal of Biogeography, 35(7), 1158-1159. https://doi.org/10.1111/j.1365-2699.2008.01928.x
Brotos, L., Thuiller, W., Araújo, M. B., & Hirzel, A. H. (2004). Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography, 27(4), 437-448. https://doi.org/10.1111/j.0906-7590.2004.03764.x
Chave, J. (2013). The problem of pattern and scale in ecology: What have we learned in 20 years? Ecology Letters, 16, 4-16. https://doi. org/10.1111/ele.12048
Chefaoui, R. M., & Lobo, J. M. (2008). Assessing the effects of pseudo-absences on predictive distribution model performance. Ecological Modelling, 210(4), 478-486. https://doi.org/10.1016/j.ecolmodel.2007.08.010
Cook, C. N., Mascia, M. B., Schwartz, M. W., Possingham, H. P., & Fuller, R. A. (2013). Achieving conservation science that bridges the knowledge-action boundary. Conservation Biology, 27(4), 669-678. https://doi.org/10.1111/cobi.12050
Cvitcanovic, C., McDonald, J., & Hobday, A. J. (2016). From science to action: Principles for undertaking environmental research that enables knowledge exchange and evidence-based decision-making. Journal of Environmental Management, 183, 864-874. https://doi.org/10.1016/j.jenvman.2016.09.038
Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. Journal of Animal Ecology, 77(4), 802-813. https://doi.org/10.1111/j.1365-2656.2008.01390.x
Evans, J. S., Murphy, M. A., Holden, Z. A., & Cushman, S. A. (2011). Modeling Species Distribution And Change Using Random Forest. In C. A. Drew, Y. F. Wiensma, & F. Huettemann (Eds.), Predictive species and habitat modeling in landscape ecology: Concepts and applications (pp. 139-159). Springer New York. https://doi.org/10.1007/978-1-4419-7390-0_8
Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., & Aalsdor, D. (2007). The shuttle radar topography mission. Reviews of Geophysics, 45(2), RG2004. https://doi.org/10.1029/2005RG000183
Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation, 24(1), 38-49. https://doi.org/10.1017/S0376892997000088
Fournace, Y., Engler, J. O., Rödder, D., & Secondi, J. (2014). Mapping species distributions with MAXENT using a geographically biased sample of presence data: A performance assessment of methods for correcting sampling bias. PLoS One, 9(5), 1-13. https://doi.org/10.1371/journal.pone.0097122
Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment, 202, 18-27. https://doi.org/10.1016/j.rse.2017.06.031
Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. Ecology Letters, 8(9), 993-1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x
Guisan, A., Thuiller, W., & Zimmermann, N. (2017). Habitat suitability and distribution models: With applications in R. Cambridge University Press. https://doi.org/10.1017/9781139028271
Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. Ecological Modelling, 135, 147-186. https://doi. org/10.1016/S0304-3800(00)00354-9
Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., & Townshend, J. R. G. (2013). High-resolution global maps of 21st-century forest cover change. Science, 342, 850-853. https://doi.org/10.1126/ science.1244693
Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology, 25, 1965-1978. https://doi.org/10.1002/joc.1276
Veloz, S. D. (2009). Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography*, 36(12), 2290-2299. https://doi.org/10.1111/j.1365-2699.2009.02174.x

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A. & NCEAS Predicting Species Distributions Working Group. (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, 14(5), 763-773. https://doi.org/10.1111/j.1472-4642.2008.00482.x

Xiong, J., Thenkabail, P. S., Gumma, M. K., Teluguntla, P., Poehnelt, J., Congalton, R. G., Yadav, K., & Thau, D. (2017). Automated crop-land mapping of continental Africa using Google Earth Engine cloud computing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 126, 225-244. https://doi.org/10.1016/j.isprsjprs.2017.01.019

Zimmermann, N. E., Edwards, T. C., Graham, C. H., Pearman, P. B., & Svenning, J.-C. (2010). New trends in species distribution modelling. *Ecography*, 33(6), 985-989. https://doi.org/10.1111/j.1600-0587.2010.06953.x

Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., Guillermo, F., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J. J., Leitão, P. J., Park, D. S., Peterson, A. T., Rapacciuolo, G., Schmatz, D. R., Schröder, B., Serra-Díaz, J. M., Thuiller, W., ... Merow, C. (2020). A standard protocol for reporting species distribution models. *Ecography*, 43(9), 1261-1277. https://doi.org/10.1111/ecog.04960

**BIOSKETCH**

**Dr. Crego** is a postdoctoral researcher at the Smithsonian Conservation Biology Institute (USA). He specializes in spatial ecology and conservation biology. He uses a combination of field data collection, remote sensing and advanced modelling to investigate how species adapt to rapid global changes.

**Dr. Stabach** is an ecologist at the Smithsonian Conservation Biology Institute (USA) and programme coordinator of the Smithsonian Institution’s Movement of Life Initiative. His research focuses on the factors that affect the abundance, distribution and movement patterns of large terrestrial mammals, incorporating remotely sensed and field data-derived variables to relate to animal point locations to address applied research questions.

**Dr. Connette** is an ecologist at the Smithsonian Conservation Biology Institute (USA). He specializes in landscape ecology, wildlife population dynamics and Geographic Information Systems (GIS). He currently serves as a science advisor for Smithsonian’s Working Land & Seascapes Initiative, which partners with communities, decision-makers and other stakeholders to inform environmental conservation policies and practices.

**Author contribution:** R.D.C and G.C. conceived of the original idea and led code development. R.D.C drafted the manuscript, with input from G.C. and J.S.

**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of the article at the publisher’s website.

**How to cite this article:** Crego, R. D., Stabach, J. A., & Connette, G. (2022). Implementation of species distribution models in Google Earth Engine. *Diversity and Distributions*, 28, 904-916. [https://doi.org/10.1111/ddi.13491](https://doi.org/10.1111/ddi.13491)