An image is worth a thousand species: combining neural networks, citizen science, and remote sensing to map biodiversity

Lauren Gillespie[1,2], Megan Ruffley[1], Moises Exposito-Alonso[1,3,4,*]

[1] Department of Plant Biology, Carnegie Institution for Science, Stanford, California, USA
[2] Department of Computer Science, Stanford University, Stanford, California, USA
[3] Department of Biology, Stanford University, Stanford, California, USA
[4] Department of Global Ecology, Carnegie Institution for Science, Stanford, California, USA

[*]to whom correspondence should be addressed: moisesexpositoalonso@gmail.com

Anthropogenic habitat destruction and climate change are altering the geographic distributions of plant communities. Although mapping vegetation changes is not possible at high resolution using remote sensing data and deep convolutional neural networks, these approaches have not been applied to model the distributions of thousands of plant species to understand spatial changes in biodiversity. To address the current lack of scalable and automatic tools to map plant species distributions at a fine-grained scale, we created a dataset of over half a million citizen science observations of 2,221 plant species across California paired with satellite images at 1 meter resolution from solely free and public sources. With this we trained a deep convolutional neural network, *deepbiosphere*, that predicts presences of plant species within 256 × 256 meter satellite images and outperforms common low-resolution species distribution models (AUC_{ROC} = 0.901, Δ = 0.026). We showcase the novelty and potential applications of this framework by visualizing high-resolution predictions of keystone species such as coastal redwoods, identifying spatio-temporal ecosystem changes from wildfires and restoration management, and detecting urban biodiversity hotspots. Deep neural networks continuously trained on public remote sensing imagery and citizen science observations could enable cheap, automatic, and scalable monitoring of biodiversity and detect rapid anthropogenic impacts on ecosystems.

Keywords: species distribution models, deep learning, remote sensing, anthropogenic change.

Human-driven land use and climate change is decreasing native plant diversity through direct extinction, fragmentation of habitats and niche shifts (1, 2). This local loss of species diversity negatively impacts ecosystem services such as carbon sequestration ability (3), primary productivity (4), local climate regulation (5), and resilience to increasingly common extreme weather events (6). To devise targeted conservation strategies of native flora and the species that depend on them, we require tools that can jointly map geographic distributions of many plant species and their micro-habitats at high resolution. Ideally, on the ground measurements of every single species present every 100m with yearly updates would provide the richest and most reliable accounting of biodiversity across ecosystems, but such approach is infeasible at scale (See **Supplementary Material [SM] 3.1**). In absence of comprehensive on the ground measurement, ultra-high-resolution remote sensing data has served as a good proxy for many important applied ecological tasks, for instance, real-time deforestation monitoring at 30-m resolution globally (7). Combining these ultra-high-resolution products with deep neural networks from computer vision has further enabled individual tree crown detection at a 1m resolution (8, 9), and vegetation type classification (10, 11). However, these methods largely focus on single keystone plant species in an ecosystem and are agnostic to the many other plants living therein, failing to capture the dynamics of the entire community. Therefore, these resulting high-resolution maps are unable to inform targeted conservation strategies at both the individual species and ecosystem scale.

Spatial mapping of many plant species is commonly done with species distribution models (SDMs). These use coarse interpolated climate maps of temperature and precipitation from weather stations across the globe to delineate regions where climate is suitable for a given species. This is based on climate thresholds inferred from the locations where such a species has been observed in the past (**SM 3**). This has enabled the systematic creation of maps of hundreds to thousands of species including maps of the majority of vascular plants native to California (bnhm-shiny.berkeley.edu/cappa) and the world (12–15) as well as maps for over two thousand of the most threatened species in the continental U.S. (16). Stacking these SDM maps of individual species together, researchers have studied general patterns of biodiversity, such as presumed changes of plant species from pre-industrial times until the 21st century (17) or projected sensitivities of species to climate change (18), and have developed large-scale biodiversity prioritization optimization models (19, 20). The application of deep convolutional neural networks for species distribution modeling across many thousands of species simultaneously has even been done from patterns in climate data with promising results (21).

However, SDMs use decadally-averaged climate maps for modeling and rarely have tapped into high-resolution remote sensing data (**SM Section 1.5,1.6, Fig. S5**) (22–24) (but see cases using topography, soil type, and land use (25, 26)). The use of interpolated climate maps as input for SDMs create coarse predictions with a maximum resolution of ~1–5 km, and miss rich information about local landscape features that determine species presence and that would be most informative for local conservation practices (**SM 1.4**). For instance, some species live on riverine environments, others are found only on north or south slopes of mountains, while others may be recent colonizations due to disturbance from human-based land use change. Given plant species often occupy fragmented landscapes affected by land use changes and oreography and species turnover can rapidly happen within decades, these coarse bioclimatic-based SDM maps cannot describe biodiversity changes from local transformations that often drive conservation consideration (**Fig. S5**), such as habitat destruction from clear cuttings, or subtle plant movements driven by climate change such as treeline advance or retreat at ecosystem edges (27, 28). To address these and better monitoring of biodiversity in the 21st century (29, 30), we envision a new generation of SDMs that recognize and map ranges of many thousands of plant species simultaneously from meters-level resolution and temporal series of remote sensing products.

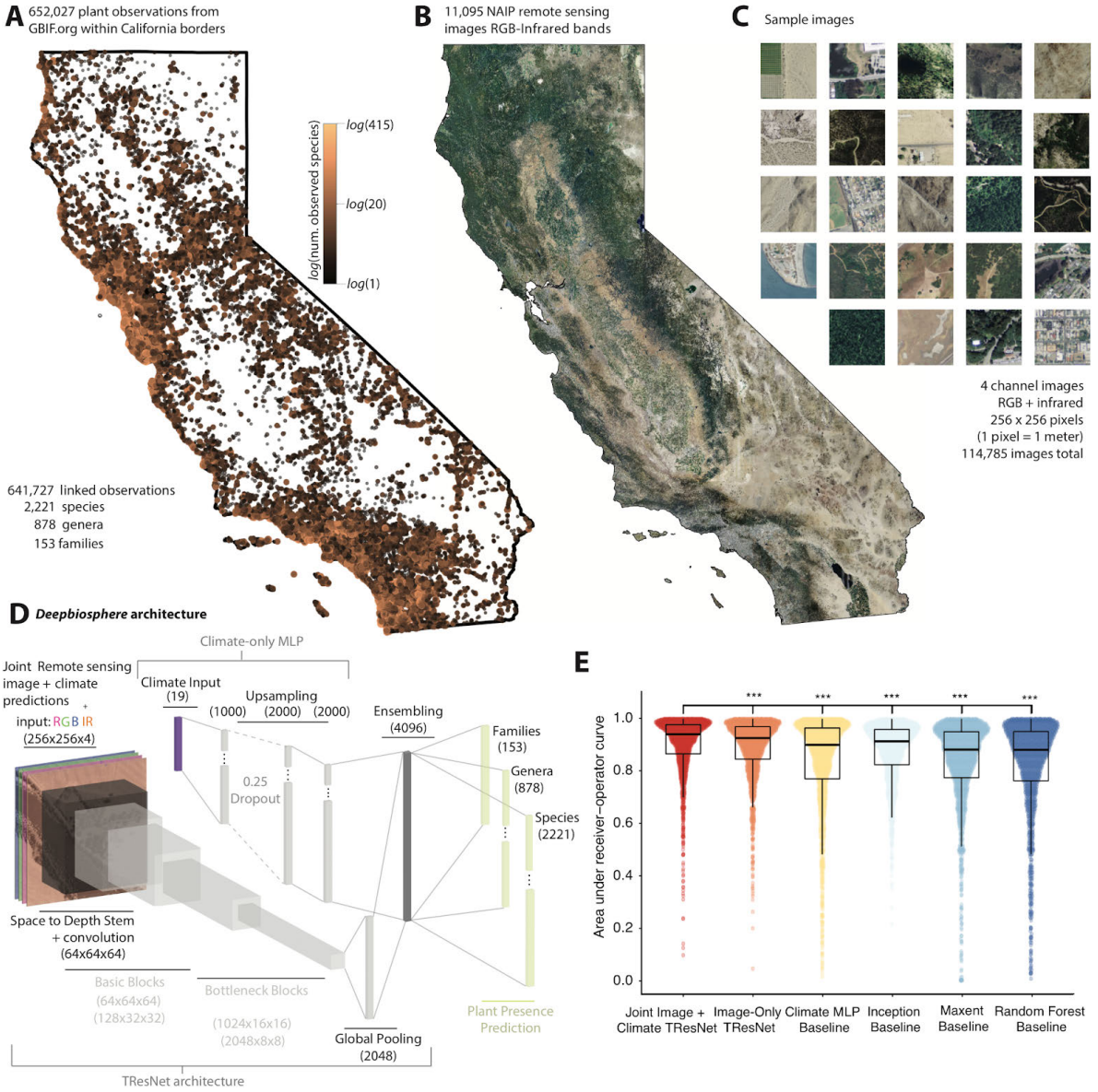

Fig. 1 | Building a multi-species distribution model (SDM) combining remote sensing imagery and citizen science data with deep convolutional neural networks. (**A**) Species observations used as labels were obtained from the Global Biodiversity Information Facility (GBIF.org). Each point is an observation from the dataset. The size and color of each point is proportional to the log of the number of species observed within 250 m at each observation point. (**B**) Aerial imagery from the National Agriculture Imagery Program (NAIP) from 2012 downloaded from the Microsoft Planetary Computer. (**C**) Sample images from the NAIP dataset. (**D**) Graphical description of the *deepbiosphere* TResNet convolutional neural network (CNN) model adapted to use climate and remote sensing imagery to predict species, genus, and family presence. The model is an online ensemble of the image-only TResNet and climate MLP models. (**E**) Comparison of the distribution of area under the receiver-operator characteristic curve (AUC_{ROC}) for each of the 2,221 species in the main dataset across SDMs. Stars indicate that AUC_{ROC} for the joint climate and remote sensing imagery TResNet are statistically significantly higher than the corresponding SDM using an unpaired student t-test with $p<10^{-5}$.

Here we test the use of convolutional neural networks (CNNs) to jointly generate SDMs of many species directly using high-resolution satellite data rather than climate to model each species' habitats (21). To do so, first we created a dataset utilizing public high-resolution RGB+infrared aerial images from 2012 acquisition of the National Agriculture Imagery Program (NAIP) (31) (**SM 1**), and combined them with public naturalist observations from the online app iNaturalist available through the Global Biodiversity Information Facility (GBIF.org). We focus on the speciose and heterogeneous region of California, where the flora has been extensively studied (**Fig. 1A**), and analyzed a total of 652,027 observations of 2,221 plant species from between the years of 2015 and 2022—nearly 30% of

the known vascular plant species of California. One byproduct of using naturalist observations is the complexity and heterogeneity of the dataset, including imbalance of observations per species, imbalance of the number of species observed by 256 × 256 m squares, or the different density of observations across geographic regions (**SM 1.7**, **Fig S2**). Most popular computer vision datasets are explicitly curated to be virtually perfectly balanced (32, 33) which makes out-of-the-box application of computer vision algorithms nontrivial (34).

For each species observation, we extracted a four channel image (red, blue, green, infrared) of 256 x 256 pixels (1 pixel = 1 meter^2) from the NAIP tiles (see **Materials & Methods**). Although most plant species will not be directly visible from this imagery, CNNs can still extract patterns of topography, land use, habitat structure, and even soil composition to provide micro-habitat descriptions complementary to climate variables. The list of species in each image was augmented to include any other species within 256 m of the focal point of an image (**Fig. S1**) with partial checklists ranging from 1 to ~400 unique species per image (**Fig. S2B**). We expect this augmentation will help CNNs learn relevant species co-occurrence relationships and potentially capture important patterns of species interactions, leading to more accurate models of plant community composition at high resolution (27, 28). Finally, we partitioned this dataset into multiple train and test splits for model testing and validation. In order to test the SDM's interpolation ability, we created a test-train split that maximized the geographic area covered by the test set (**SM 1.3.1**, **Fig. S6**). To test the SDM's extrapolation ability and properly validate the model's performance, we also used a 5-fold spatial cross-validation approach (**SM 1.3.2, Fig. S7**). Finally, we utilized a wide variety of accuracy metrics from across a variety of relevant disciplines (**SM 2**). This novel dataset of over half a million images and thousands of species was created with a completely free, open-source pipeline that automatically downloads, cleans, and collates naturalist observations, along with other early efforts (25), should serve as a benchmark for the machine learning community to test models connecting biodiversity and satellite data (code available at github.com/moiexpositoalonsolab/deepbiosphere).

Our chosen CNN architecture was the TResNet architecture (35) which is optimized for uneven multilabel image classification and presence-only data. The TResNet architecture has been tested to classify 1,000 every-day human objects (ImageNet) and achieves human-level accuracy (**Fig. S8**, **SM 3.2, Table S2**). We initially trained the TResNet with only the NAIP aerial imagery, hereafter referred to as *image-only TResNet* architecture. We also modified the TResNet architecture to use both NAIP images and climate data as input, referred to as the *image + climate TResNet* architecture (**Fig. 1D, Table S2**). After multiple attempts to account for the biases present in the dataset (**SM 2.7, Fig. S3**), we developed a novel loss function, a frequency-scaled binary cross entropy (Scaled BCE) which outperforms common loss functions (**Table S3**) when paired with the image-only TResNet architecture, as the network maximizes learning from those species present rather than absent (note that absence in the data can be due to lack of observation efforts, see **Materials & Methods, SM 3.2.1**). We compared the two TResNet architectures with the Scaled BCE loss function against other climate-based SDM models and deep neural network architectures (**Table 1, Table S3**), including standard SDMs like MaxEnt, random forest, and multi-perceptron neural network (MLP) with standard WorldClim climate variables as inputs (24) (**SM 3.3**, **Table S2**). We also compared the TResNet architectures to an Inception V3 CNN model that ranks species presence from NAIP images (winner of the GeoLifeCLEF 2020 challenge (34), see **SM 3.2.5**, **Table S2**). Finally, we also report an image-only TResNet trained using a smaller, preliminary dataset with only 100,000 images which we refer to as $deepbiosphere_{alpha}$ (**SM 3.2.4, Table S3**).

Table 1 | Accuracy comparison among tested CNNs and baselines on interpolation split of dataset. Mean and standard deviation of accuracy metrics per-species ($_{spp}$ subscript) or per-image ($_{Image}$ subscript) in interpolation test set are reported for each species distribution model as well as basic random or frequency-based baselines. AUC_{ROC} = area under receiver operator curve; Cal. AUC_{ROC} = calibrated area under the receiver operator characteristic curve; AUC_{PRC} = area under precision-recall curve; Rec= recall; Prec = precision; Top 30= top 30 accuracy. Climate-based models are built using bioclim variables which come at 30 arc-second resolution, which translates to approximately 1,000-1,300m sized pixels within California's latitudes. The image+climate TResNet (img+clim TResNet), image-only TResNet (img TResNet), and MLP (climate MLP) models were trained with a learning rate of 0.0001 with a batch size of 65 and 1,000, respectively. The Inception baseline used a batch size of 100 and a learning rate of 0.01 (34). Random forest and MaxEnt SDMs were trained individually for each species and aggregated across species with 50,000 background samples following best practices (36). The frequency baseline involved calculating the frequency of observations per-species on the training set, rescaling the probabilities to 0.001-1.0 and imputing these frequencies as the predicted probabilities at each test set example. The random baseline was calculated by drawing random values from a standard normal distribution ten times and averaging the accuracy metrics across these ten trials.

Model name	Resolution	Loss	$AUC_{ROC\ spp}$	Cal AUC_{ROC}	$AUC_{PRC\ spp}$	$Prec_{Image}$	$Prec_{Spp}$	$F1_{Spp}$	Top 30_{spp}
img-clim TResNet	256 m	Scaled BCE	**0.901±0.117**	**0.014**	**0.086±0.133**	**0.072±0.106**	**0.057±0.102**	**0.081±0.119**	**0.263±0.305**
img TResNet	256 m	Scaled BCE	0.887±0.121	0.01	0.073±0.117	0.058±0.091	0.043±0.075	0.067±0.096	0.248±0.299
Climate MLP	~1,000 m	Scaled BCE	0.83±0.193	0.007	0.075±0.127	0.058±0.088	0.038±0.077	0.061±0.104	0.208±0.27
Inception (34)	256 m	CE	0.875±0.116	0.002	0.04±0.073	0.0±0.0	0.0±0.0	0.0±0.0	0.208±0.307
Random Forest	~1,000 m	N/A	0.817±0.196	0.008	0.084±0.142	0.029±0.051	0.041±0.077	0.065±0.109	0.171±0.274
MaxEnt	~1,000 m	N/A	0.829±0.177	0.008	0.073±0.129	0.021±0.041	0.041±0.088	0.063±0.122	0.081±0.196
Frequency	NA	N/A	0.5±0.0	0.002	0.005±0.008	0.039±0.116	0.0±0.005	0.001±0.009	0.019±0.138
Random	NA	N/A	0.499±0.036	0.002	0.005±0.008	0.005±0.01	0.005±0.008	0.009±0.015	0.014±0.024

The image+climate TResNet model trained with the novel Scaled BCE function was the best performing model overall with the highest average area under the receiver-operator curve, among other metrics (AUC_{ROC}= 0.901±0.117, **Fig. 1E**, per-species recall=0.278±0.317, per-species precision= 0.057±0.102, per-species F1=0.081±0.119, see **Table 1, Figs. S9-12**). For the majority of metrics, this model had superior performance across all types of accuracy metrics, including ranking metrics like top 30 accuracy per-species, and binary classification metrics, like per-species F1. It was also superior to its TResNet counterpart trained only with images (**Table 1**) or with other loss functions (Δ per-species F1 =1.4-7.2% improvement, Δ AUC_{ROC}=2.6-10% improvement, Table S3), and the CNN baseline Inception model (**Table 1**). It also increases prediction accuracy compared to common climate-based SDMs like random forest, MaxEnt, and non-convolutional neural network MLP and . The image+climate TResNet model trained with Scaled BCE loss also performed well in spatial cross-validation permutations of the dataset (**Table 2**), with the model exhibiting higher accuracy on all metrics except recall (Table S4). This top model—hereafter called *deepbiosphere*—therefore not only predicts species distributions at much higher spatial resolution than common SDMs (up to 1 m^2 theoretical resolution from remote sensing imagery), but it appears to be extracting important information from colors and textures in the landscape, likely related to topographic and/or biotic factors thought to influence species' distributions (27, 28).

Table 2 | Accuracy comparison among tested CNNs and baselines. Mean and standard deviation are reported across the five cross-validation bands for each species distribution model. AUC_{ROC} = area under receiver operator curve; Cal. AUC_{ROC} = calibrated area under the receiver operator characteristic curve; AUC_{PRC} = area under precision-recall curve; Rec= recall; Prec = precision; Top 30= top 30 accuracy. Climate-based models are built using bioclim variables. The image-only TResNet (img TResNet) and MLP (climate MLP) models were trained with a learning rate of 0.0001 with a batch size of 65 and 1,000, respectively using Scaled BCE loss. Random forest and MaxEnt SDMs were trained individually for each species and aggregated across species with 50,000 background samples following best practices (36).

Model name	Resolution	Loss	$AUC_{ROC\ spp}$	$AUC_{PRC\ spp}$	Cal. AUC_{ROc}	Rec_{Spp}	$Prec_{Spp}$	$F1_{Spp}$	Top 30_{spp}
img-clim TResNet	256 m	Scaled BCE	**0.069±0.013**	**0.803±0.026**	**0.028±0.006**	0.136±0.022	**0.068±0.016**	**0.075±0.016**	**0.13±0.019**
Climate MLP	~1,000 m	Scaled BCE	0.054±0.009	0.693±0.03	0.019±0.004	0.127±0.017	0.044±0.01	0.056±0.01	0.087±0.014
Random Forest	~1,000 m	N/A	0.057±0.01	0.708±0.019	0.011±0.002	**0.485±0.129**	0.041±0.01	0.063±0.018	0.087±0.009

Model name	Resolution	Loss	$AUC_{ROC\ spp}$	$AUC_{PRC\ spp}$	Cal. AUC_{ROc}	Rec_{Spp}	$Prec_{Spp}$	$F1_{Spp}$	Top 30_{spp}
img-clim TResNet	256 m	Scaled BCE	**0.069±0.013**	**0.803±0.026**	**0.028±0.006**	0.136±0.022	**0.068±0.016**	**0.075±0.016**	**0.13±0.019**
MaxEnt	~1,000 m	N/A	0.057±0.013	0.722±0.015	0.011±0.002	0.396±0.149	0.039±0.011	0.06±0.019	0.056±0.013

To showcase the insights gained from *deepbiosphere* beyond accuracy metrics, we created several case studies where species prediction outputs of *deepbiosphere* can be compared with known biodiversity patterns across California (**SM 4.1**). The training of *deepbiosphere* on 2,221 species included 77,915 coastal redwood observations by citizen naturalists. Extracting the probability output of our model corresponding to *S. sempervirens* around Prairie Creek in Redwoods National Park we can verify that it distinguishes groves of old-growth trees from other evergreen forests without redwoods (**Fig. 2A-D**, AUC_{ROC}=0.979, Recall=0.463, Precision=0.281, F1=0.35) and predictions agree with manual human-annotated cover (Pearson's r = 0.728, P-value < 0.00005 with Dutilleul test, cohen's kappa =0.251, **Fig. 2D, S15; SM 4.1.1**). Expert-designed MaxEnt models from Baldwin et. al. ((19), https://bnhm-shiny.berkeley.edu/cappa) predict the probability of redwood presence = 0.4497 (often not considered present under the common 0.5 threshold) for the entire 5 x 8 km^2 Prairie Creek region. Additionally, the MaxEnt baseline trained on the same observations as *deepbiosphere* show no correlation with human annotations (Pearson's r_{MaxEnt}: -0.092, P-value = 0.636 Dutilleul test; **Fig. S16**) and does not correctly predict the known redwood occurrences from the training set (**Fig. S16B**), while the Inception baseline shows an inverse relationship with the human annotations when using the correct sigmoid transformation (Pearson's r = -0.46, P-value < $4x10^{-5}$ with Dutilleul test, **Fig. S17**). Importantly, because both expert-drawn and our own MaxEnt SDM are fitted with bioclimatic variables, the prediction maps are at 1 km resolution and naturally are unable to capture the habitat fragmentation of the remaining old-growth forest (**Fig 2B, Fig S18**).

Not only large species of plants can be well-predicted by *deepbiosphere*. The understory herb redwood sorrel (*Oxalis oregana*), often associates more with old-growth redwood habitats, and the shrub red huckleberry (*Vaccinium parvifolium*), which tends to associate more with post-clearcutting secondary growth (37), are also well predicted ($AUC_{ROC\ Oxa}$ = 0.990, $AUC_{ROC\ Vacc}$ = 0.982) (**Fig. S19B,C**) but more interestingly *deepbiosphere* visually appears to be disambiguating that redwood sorrel more readily associates with the old-growth redwoods while red huckleberry prefers secondary growth (**Fig. S19D**). *Deepbiosphere*'s ability to capture subtle relationships between important indicator species also extends to the oak savannas of Southern California, where it can delineate transitions between rapidly changing, interwoven habitats in the arid Santa Ynez mountains (**Figs. S20-22, SM 4.1.3**). These examples strengthen the idea that combining remote sensing deep learning and citizen science observations could potentially be used in a variety of monitoring and conservation tasks such as measuring habitat fragmentation or disambiguating vegetation alliances.

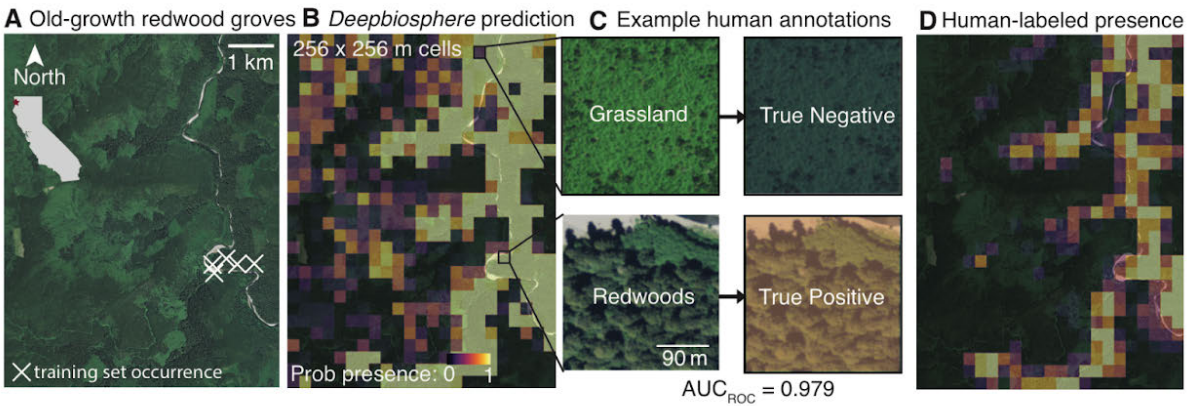

Fig. 2 | Proof-of-concept detection of redwood forest (*Sequoia sempervirens*) at a high resolution from a *deepbiosphere*'s network trained on 2,221 plant species. (**A**) Location of case study in Redwoods National Forest in California and example NAIP imagery, along with five redwood occurrences used in training models. (**B**) The same

geographic coordinates divided into 256 × 256m squares with overlaid probability of presence from *deepbiosphere*. (**C**) Zoom-in examples of either grassland or redwood annotation. (**D**) Human annotation of old-growth redwood presence.

We also wondered whether *deepbiosphere* could also be helpful to get insights of coordinated changes in plant community composition across space and time, whether these are local hotspots of high numbers of species, transitions in edges between two ecosystems, or ecological successions or transitions over time. For this analysis, we utilized $deepbiosphere_{alpha}$, an image-only TResNet trained using a smaller, preliminary dataset but also of a large number of 2,242 species (**Table S1**). We first used the output vector of $deepbiosphere_{alpha}$ predictions of presences across all species to map areas of apparent high plant biodiversity. We considered all species with predicted probability >50% as present in each 256 × 256m pixel across the San Francisco metropolitan area (**Fig 3B).** Visually, the green areas of San Francisco, such as Golden Gate Park and Angel Island, appear to have significantly higher predicted plant biodiversity. To quantify this observation, we classified each pixel in San Francisco as either park or urban area depending on if it fell into the outline of any registered park in Marin, San Francisco, and San Mateo counties (**SM 4.2.1, Fig. S23**) and found that pixels in parks had significantly higher predicted biodiversity than pixels in urban areas (unpaired student's t-test, $p < 2.2 \times 10^{-16}$, **Fig. 3D**). This predicted biodiversity correlates with the density of dataset observations from iNaturalist ($R = 0.42$, P-value $< 10^{-5}$ with Dutilleul correction) (**Fig. S24C**), and naturally reflects the algorithm learned to make predictions in areas that has *seen* before. However, species richness predictions and number of training observations is far from perfect (**Fig. S24**) and the algorithm highlights areas of potentially high diversity with low comparable naturalist observation density (south San Francisco area) as well as areas of high observation density but low diversity (downtown San Francisco piers). This could be potentially helpful to automatically and quantitatively map the value of green hotspots within urban areas and plan for the impacts of potential new developments in cities.

We were also curious what patterns in ecosystem transition $deepbiosphere_{alpha}$ could also passively uncover. Taking inspiration from edge detection algorithms common in computer vision image processing, we used a one-neighbor convolution (i.e. looking at the 8 cells surrounding a focal cell) to calculate the average L^2 norm of the sigmoid-transformed probabilities of neighboring cells from the current cell (**SM. 4.2.2, Fig. S25**). Interpreting this as a proxy metric for the amount of local spatial turnover within a 512m radius, we sought to identify borders of high species turnover. We chose to focus on Point Reyes (**Fig. S26A**), where the coast-forest and forest-grassland transitions are obvious and where independent manually-curated datasets of vegetation transitions are available from the county administration (38). We found that the intensity of the predicted spatial turnover correlated with the number of curated vegetation classes within 256 x 256-m pixels (Pearson's $r = 0.391$, P-value $< 10^{-5}$; **Fig. S69D**) and did not correlate with the observation density (Pearson's $r = 0.0597$, P-value = 0.014 with Dutilleul test) (**Fig S27**). Transitions between ecosystems, so-called ecotones, are typically of high biodiversity value and can be important to delimit protection areas, thus supporting the utility of predicting species distributions for entire assemblages of species rather than single species, as often conducted in SDMs.

Finally, we asked whether *deepbiosphere* could detect temporal shifts in plant communities (**SM 4.3**). To look at successional community shifts, we compared the predicted species probabilities in the Hetch Hetchy Valley near Yosemite National Park between 2012 and 2014, bracketing the 2013 Rim Fire that burned over 250,000 acres (**Fig. 3A**). Specifically we calculated a metric of community change by first making predictions of species presence applying $deepbiosphere_{alpha}$ to NAIP satellite images taken in 2012 and 2014, and then extracting the Euclidean distances between species probabilities predicted before and after those timepoints (an example can be seen in **Fig. S28, SM 4.3.1**). As expected, community change as estimated from Euclidean distance was elevated within the burn scar (**Fig. S29**) and correlated strongly with an empirical metric of burn severity—normalized differential burn ratio—commonly used in fire ecology and quantified by Oak Ridge National Laboratory (39) using hyperspectral MASTER data validated with field estimates, (Pearson's $r = 0.54$,

P-value < 10^{-4}; **Fig. 3C**). On an individual species level, the change in predictions from *deepbiosphere*$_{alpha}$ also follow ecological expectations for some charismatic species after the fire, with a decreased presence of species like Aspen whose adapted strategy to fire is for mature trees to burn and regenerate from the root system after the fire years later (**Fig. S30**) and an increased presence of some fire-following herbs like Blue Dicks (**Fig. S31**). Temporal trends appear not only obvious in negative impacts such as forest fires, but also positive impacts, for instance a recent wetlands restoration project in northeastern California at the Ash Creek Wildlife Refuge (40) (**SM 4.3.2**, **Fig. S35, S36**).

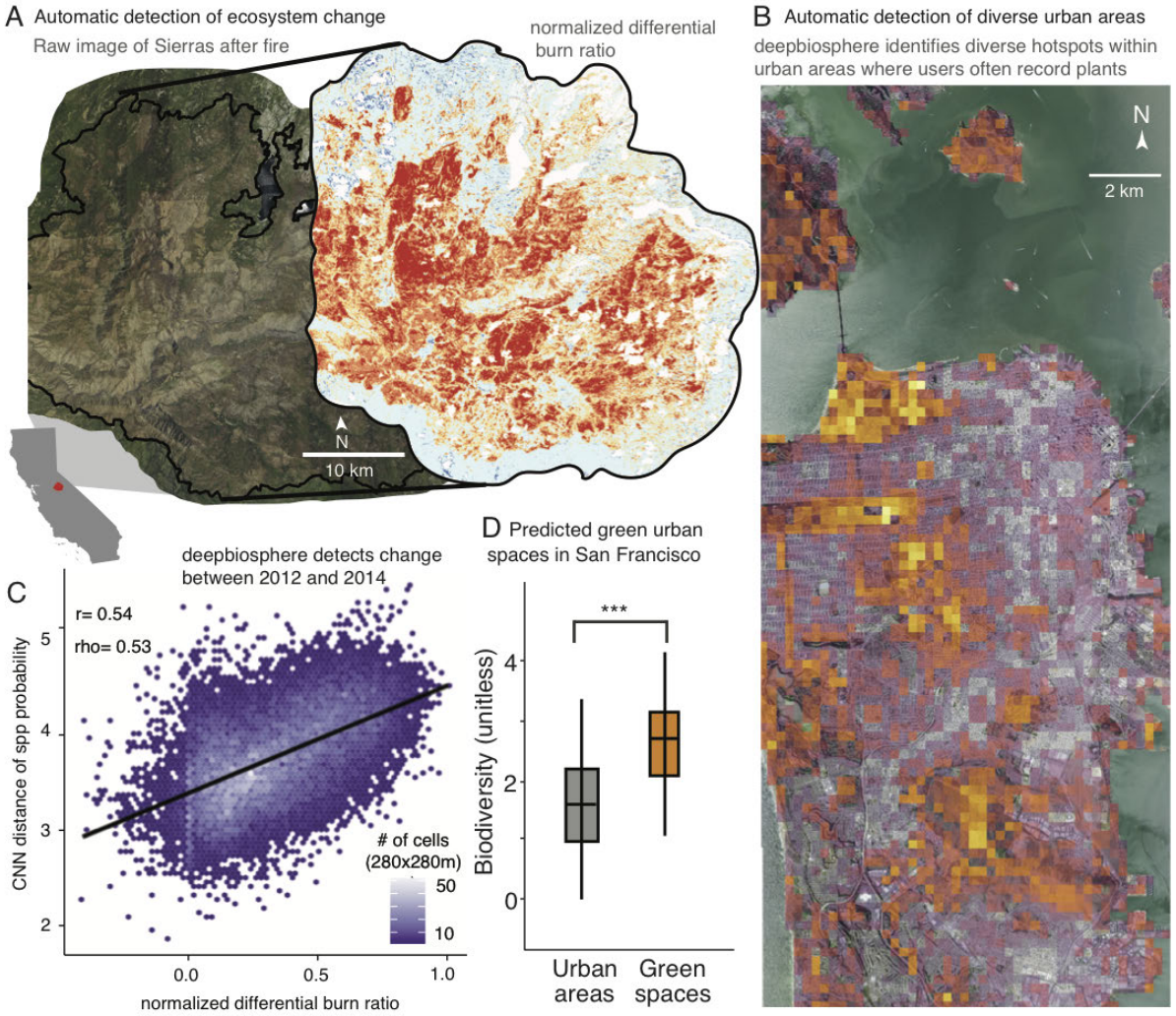

Fig. 3 | Potential application of *deepbiosphere*'s species predictions to detect ecological changes across California. (**A**) Example NAIP aerial image from after the 2013 Rim Fire in the Sierra mountains. (**B**) Detection of highly diverse urban green spots in San Francisco. (**C**) Comparison of *deepbiosphere*$_{alpha}$*'s* predicted community change to empirical methods of measuring burn severity. (**D**) Comparison of predicted local species diversity for 256 × 256 m pixels that map to either green or urban space. Comparison was done pixel by pixel and determined to be statistically significant using an unpaired student t test (*** indicates a P-value < 10^{-4}).

Despite the better accuracy performance than traditional SDMs and the potential applications in mapping or detecting change in biodiversity, absolute performance metrics of *deepbiosphere* are lower than the often-reported in machine learning literature (41, 42). Despite that this is one of the largest dataset available of paired satellite images and plant species checklists, it barely approaches the large number of images from state-of-the-art benchmarks in deep learning (32), but given the current trends in citizen science engagement and satellite-based Earth monitoring missions, we only expect these datasets to grow exponentially. Although here we report many lessons learned from utilizing different

CNN architectures, loss functions, and training techniques, further efforts are also needed to develop new methods to overcome real-world species observation biases and expand validation through curated ground truth species lists. While we do not expect that *deepbiosphere* will be able to predict all species at all locations well on average, due to limited numbers of examples for many species and dataset imbalances, we do expect the network to have learned the habitat of a good number of well-sampled species and could serve as an early indicator of biodiversity changes.

Deepbiosphere is a first proof-of-principle multi-species species distribution model based on convolutional neural networks and high-resolution satellite data. Trained with weak labels of species observations by citizen science, this model enables automatic and accurate fine-scale mapping of many species, and could potentially provide early warnings of habitat impacts in urban centers or natural ecosystem transformations from ecological disasters. *Deepbiosphere*'s deep learning approach represents a much called-for paradigm and philosophical shift of modeling species distributions (29, 30, 43), where trained models could readily be coupled with continuously updated public satellite and citizen science species observations to generate near-real-time updates of biodiversity change. Modern advances in deep learning and cloud computing should eventually enable scaling the presented framework to a continental or global scale. We envision deep learning will enable better modeling plant species distributions at high spatio-temporal resolution and help international policies to define protection areas and preservation of biodiversity.

METHODS SUMMARY

Species observations. Inspired by the GeoLifeCLEF dataset (25), we collected observations from kingdom *Plantae* using *GBIF.org* (44) from the years 2015-2022. Only records observed by humans with a coordinate uncertainty radius of less than or equal to 120m with no flagged geospatial issues were taken from within the state of California. We downloaded a total of 912,380 observations of 5,193 unique plant species which were further filtered to only include vascular plants, remove duplicate observations of the same species within a 150m radius, remove species that contain all observations located within a 256m radius, were not geographically located within the Global Administrative Area boundary of California (45), and were not located within both the climatic and remote sensing imagery rasters. To increase the density of observations in the dataset, we used neighbor imputation to add any other species observed within an overlapping 256m radius to a given observation (**Fig. S1**). We finally removed any species that had fewer than 500 total observations in the dataset after neighbor imputation, leaving us with a total of 652,027 observations of 2,221 unique plant species (**Fig. 1A, Table S1**).

Remote sensing data. To create images for each observation, we utilized aerial imagery from the National Agricultural Imagery Program (NAIP) (31) which we downloaded for the entire state of California from 2012 and 2014 using Microsoft Azure's NAIP data blob disk image on its West Europe and Eastern U.S. servers. For training the CNN models, we specifically used the NAIP data from 2012 (31) at 1-m resolution to generate 256 x 256 pixel images, where 1 pixel corresponds to a 1 x 1 meter resolution. We used all available bands, specifically a three-color band (RGB) and infrared. While other missions such as Landsat include multi-spectral information, we believe that both higher resolution and simpler image products will enable scalability of methods to regions in the world with less available imagery and are better fitted for the task of plant species identification (**Fig. S3**). To make the image dataset, for each species observation we extracted a 256 × 256 pixel image from this imagery in all four RGB-Infrared bands with the central pixel at the location of the observation.

Climate variables. For climate data, we used the WorldClim Version 2 bioclimatic variables at 30 arc-second (approximately 1km) per-pixel resolution (24). Variables were downloaded directly from the WorldClim repository (http://www.worldclim.com/version2). Before fitting any model, all bioclimate variables were normalized per-variable to mean 0 and standard deviation of 1 using the entire raster clipped to the outline of California.

Splitting data for model validation. In order to properly validate and compare models, we split the dataset into multiple partitions. The first partition, which was used for hyperparameter tuning and loss comparison, was

generated by uniformly randomly selecting observations from across the state to be part of the test set and to test interpolation ability (**Fig. S6**). We chose points uniformly across the state to maximize the number of unique environments models would be evaluated on. To ensure the independence of training and testing set data due to spatial autocorrelation, we added all overlapping observations to the test set to guarantee that none of the remote sensing images and observations in the test set were present in the training set. In order to ensure that there was no data leakage between the test and train set only observations which were more than 1,300m away from any other non-overlapping observation were included. We chose an exclusion radius of 1,300m because the climate variable raster pixels converted from arc-seconds to meters can have a diameter of up to 1,200m, so any test set observation within that distance to any observation in the train set would have an identical input value as some observations used during fitting, also known as data leakage. Ultimately 1.88% of the dataset was set aside for testing.

In order to provide cross-validation of the uniform train-test split and to test the extrapolation ability of all models, we also conducted a 10-fold spatial holdout block validation by partitioning California into five 1 degree latitudinal bands (**SM Fig. S7**). Train set points within 1,300m of the test band were removed to prevent data leakage as discussed above. For models utilizing pseudo absence points, all pseudo absence points within the test bands were removed to ensure a fair comparison to presence-only models. Ultimately, the percentage of test points per-spatial block ranged from 1.40-25.35% of the entire dataset.

Convolutional Deep Neural Network definition. We implemented our Convolutional Neural Networks (CNN) using the Pytorch framework (https://pytorch.org) (46). For our final architecture, we chose the TResNet architecture of medium size (35). We chose this specific CNN architecture as it is a state-of-the-art architecture for multi-label image classification on the popular MS-COCO and NUS-WIDE datasets, two standard benchmarks for multi-label image classification in the computer vision community. We slightly modified the default TResNet architecture to have four input channels in order to support the infrared band and three fully connected output layers corresponding to the three taxonomic ranks being predicted: family, genus, and species, which we refer to as the image-only TResNet architecture (**Fig. S8**, **Table S2**).

We also modified the TResNet architecture further to include a fully connected MLP with 0.25 dropout and eLU activation function inspired by Battey et. al. (47) whose predictions were concatenated with predictions from the image-only TResNet before passing through one final fully-connected layer with ReLU activation function the three fully-connected output layers (**Fig 1D, Table S2**). We refer to this architecture as the image+climate TResNet. For analysis, all model outputs were converted to independent probabilities using the sigmoid function. For comparison to previous work using CNNs to rank species presence from remote sensing imagery (34), we trained an Inception V3 architecture with softmax cross entropy loss using the official implementation from pytorch and using both the standard and auxiliary loss during training (48). We utilize the standard dropout rate of 0.5 and a standard learning rate of 0.01, different but comparable values to those used in previous work (34).

Weights were initialized following best practices laid out in the original TResNet paper, using Kaiming He-style for CNN layers and zeroed out BatchNorm and residual connections (35). We compared performance of the image-only TResNet model on a variety of loss functions, which is discussed in further detail in **SM 3.1.1**. We found the best loss function to be a novel loss function we refer to as scaled binary cross-entropy loss (scaled BCE) which simply weights the loss contribution of present versus absent classes by the number of present and absent classes, respectively. The effect is that the magnitude of the contribution to the loss of present labels versus absent labels is approximately equal; in other words the correctness of the predictions for present species matters as much as the correctness of the predictions for absent species when calculating the model's regret. The model referred to as *deepbiosphere* is trained using this loss.

$$\text{Scaled BCE} = -L_+ - L_- \begin{cases} L_+ = \frac{\bar{y} \cdot log(f(\hat{y}))}{\sum_1^S \bar{y}_i} \\ L_- = \frac{(1-\bar{y}) \cdot log(1-f(\hat{y}))}{\sum_1^S 1-\bar{y}_i} \end{cases} \quad \text{where} f(\hat{y}_i) = \frac{1}{1+e^{-\hat{y}_i}}$$

We also report accuracy and perform downstream analysis using an image-only TResNet trained using a preliminary version of the dataset and trained with the asymmetric focal loss function.

$$\text{ASL} = -L_+ - L_- \begin{cases} L_+ = \bar{y} \cdot (1 - f(\hat{y}))^{\gamma+} \cdot log(f(\hat{y})) \\ L_- = (1 - \bar{y}) \cdot p(f(\hat{y}))^{\gamma-} \cdot log(1 - p(f(\hat{y}))) \end{cases}$$

$$\text{where } f(\hat{y}_i) = \frac{1}{1 + e^{-\hat{y}_i}} \text{ and where } p(\hat{y}) = max(p - m, 0)$$

All neural network models were trained with standard mini-batch stochastic gradient descent using the Adam optimizer (49). Batch size, total number of epochs, memory usage and training time can be found in **Table S3**.

Species Distribution Models based on climate rasters. There are a wide variety of popular SDMs, ranging from simple regression to ensembles of dozens of models, but the types of tasks that SDMs were used for tend to be very specific for either a given region or given taxon of species. Use cases range from assessing species invasion to determining conservation priority, or detecting possible new populations of a rare species (50). However, in order to generate high-quality SDMs using traditional methods, many involved steps and decisions must be made (see Table 2 of (50) for specifics) including hand collection of occurrence records by practitioners or thorough cleaning of public occurrence records, careful choice in environmental correlates, and in-field validation of predictions to ensure that predictions are logical and well-supported. Such rigorousness in modeling as spelled out in (50) is still considered best practices for modeling individual species or modeling within specific geographic regions, and when using proper data preparation and validation methods it's doubtful that any automated approach will replace expert, hand-designed SDMs for targeted modeling tasks that are narrow in scope. However, the diligence necessary to generate such high-quality expert-designed SDMs is difficult to scale to the level of the hundreds of thousands of known plant species across the world, and even within California expert-drawn SDMs for 5,500 plant species still do not manage to follow all the aforementioned best practices (51).

In light of these limitations of scale and resources, in this work we attempt to take a different, more automated approach to species distribution modeling where instead of focusing on a specific target taxa or task we rather attempt to encompass as many taxa and geographic regions as possible with as few priors as possible on what potential tasks the model could be used for. We also attempt to minimize the amount of data cleaning and taxa-specific environmental correlates preparation that is usually necessary to build a high-quality SDM. Finally, instead of attempting to fit a model to answer a specific ecological question or help solve a specific conservation task *a-priori*, we instead take a *post-facto* approach and instead observe what interesting ecological phenomena emerge from the model's predictions, rather than trying to model those phenomena explicitly from the outset.

To this note, when generating traditional bioclimatic SDMs to compare our CNN-based approach against as baselines, we intentionally chose to ignore best practices steps that require manual curation. Therefore, we chose to use default settings when possible and only performed data cleaning steps that required just the underlying dataset. We feel this "simplicity first" approach for traditional SDMs is not just a fairer comparison to the CNN-based approaches but also tracks closer to what a new SDM users and practitioners may perform. Concretely, we use the popular *dismo* package for species distribution modeling (52) and compare against two popular SDM approaches: MaxEnt and downsampled single stacked random forest using best practices lined out in (36). We chose these two models specifically from the dozens of approaches tried in (36) as these two models had consistently the best performance across the hundreds of species in their dataset, besides ensembling approaches. We also attempted to compare against ensembling approaches and run the popular *biomod* ensembling algorithm, but the algorithm was too slow and memory-intensive for us to be able to run it on all species in our dataset.

We used bioclim variables normalized to mean 0 standard deviation 1 as outlined above, and consistent with (36) we removed all but one bioclim variable with a pearson correlation coefficient higher than 0.8, leaving ten variables in total for modeling including Mean Diurnal Range, Max Temperature of Warmest Month, Minimum Temperature of Coldest Month, Annual Precipitation, Precipitation of Wettest Month, Precipitation of Driest Month, Precipitation Seasonality, Precipitation of Wettest Quarter, Precipitation of Warmest Quarter, and Precipitation of Coldest Quarter.

For each species, we generated 50,000 background samples consistent with (36) by generating a circular overlay across all points in the training dataset where the radius of each circle is the median distance between observations in the dataset. For the extrapolation experiments using the spatial cross-validation approach outlined above we removed all background samples within the spatially withheld portion of the state. Finally, we

used the same background points for both the random forest and MaxEnt models to ensure a proper comparison. For running both models, we utilized the same settings as used in (36), namely the "nothreshold" option for MaxEnt and 1,000 trees with equal bootstrapping of positive and negative samples with replacement. All other options were the default settings that come in *dismo*.

Finally, we also trained a four layer fully-connected, feed-forward multilayered perceptron (MLP) on all 19 bioclim variables as a climate-only deep learning baseline. The model's architecture was inspired by (47) and consisted of two fully-connected layers with 1,000 neurons each, followed by a dropout layer with a 0.25 dropout rate, then by two layers with 2,000 neurons each, before predicting species, genus, and family (**Table S2**). Batch size, total number of epochs, memory usage and training time can be found in **Table S2**.

Accuracy metrics. Many accuracy metrics exist for measuring the reliability and utility of both SDMs and computer vision CNN models. Accuracy metrics are much like SDMs in that no metric is perfect and able to describe all relevant and interesting behaviors a model may exhibit in one single value. Rather, each metric will highlight a specific ability of a model, such as its discrimination ability, its calibration, among other, and oftentimes, different communities will use and value different metrics. For example, popular tasks in the computer vision community tend to focus on detecting or labeling a specific object or label and thus the community cares most whether the highest predicted labels are the correct one, thus focusing heavily on models' ranking ability. This means computer vision papers tend to report top-K accuracy metrics, metrics that are most useful for tasks that are single-label in nature. On the other hand, many SDMs are used for habitat suitability modeling and thus the calibration and discrimination of the models' prediction of presence are extremely important, meaning the community prefers to report the area under the precision-recall or receiver-operator curve, averaged over species.

Because different communities are calibrated to different metrics, which accuracy metrics to use and to focus on become a non-trivial question, and judgment of which model is the "best" may vary across communities. To that note, we choose to report eighteen accuracy metrics spanning metrics common to fields from species distribution modeling to computer vision to document retrieval and ranking. Some metrics are more relevant to our multi-labeled task of joint species distribution modeling and thus in the interest of space in the main paper we report only those metrics we consider most relevant, but we report all metrics in the supplemental for completeness and openness about all models' strengths and weaknesses. For a more complete description of each accuracy metric and discussion on what skills each metric highlights, refer to the supplemental (**SM 2**). We calculated accuracy metrics only for species present in both the train and test split.

ADDITIONAL INFORMATION

Data availability. Data is publicly available through GBIF.org and NAIP (31). Scripts to regenerate paired image-species datasets and to build *deepbiosphere* are available at github.com/moiexpositoalonsolab/deepbiosphere.

Author contribution. M.E.-A. conceived the project. M.E.-A., M.R. and L.G. developed scripts and conducted the research. L.G. prepared the first manuscript draft. M.E.-A., M.R. and L.G. discussed and edited the final manuscript.

Acknowledgements. We thank Mateo Rojas-Carulla, Carl Johann Simon-Gabriel, and Detlef Weigel for early discussions or support on this project. We thank Avery Hill, the Moi Lab and Goodman Lab for comments and discussion. We also thank Claudia Engel, Annette Jing, and Sifan Liu for assistance with spatial statistics. Furthermore, we thank Lucas Czech, Gabriel Poesia Reis e Silva, Shannon Hately, Jason Thomas, and Catrina Gillespie for help with annotations. Finally, we thank Ken Morefield and Rosalie Yacoub for assistance with acquiring data.

Funding statement. This research was funded by the Carnegie Institution for Science and by Microsoft Azure AI for Earth compute credits grant (M.E.-A.). This research was also funded by the NSF Graduate Research Fellowship DGE-1656518 and the Stanford TomKat Fellowship. Part of the

computing for this project was performed on the Memex, Calc, and MoiNode clusters of the Carnegie Institution for Science and the Caltech Resnick High Performance Computing Center.

Disclosure statement. The authors declare no competing financial interests. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

1. T. Newbold, *et al.*, Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. *Science* **353**, 288–291 (2016).

2. C. Parmesan, G. Yohe, A globally coherent fingerprint of climate change impacts across natural systems. *Nature* **421**, 37–42 (2003).

3. Y. Yang, D. Tilman, G. Furey, C. Lehman, Soil carbon sequestration accelerated by restoration of grassland biodiversity. *Nat. Commun.* **10**, 718 (2019).

4. J. Liang, *et al.*, Positive biodiversity-productivity relationship predominant in global forests. *Science* **354**, aaf8957 (2016).

5. A. T. Leite-Filho, B. S. Soares-Filho, J. L. Davis, G. M. Abrahão, J. Börner, Deforestation reduces rainfall and agricultural revenues in the Brazilian Amazon. *Nat. Commun.* **12**, 2591 (2021).

6. B. Sakschewski, *et al.*, Resilience of Amazon forests emerges from plant trait diversity. *Nat. Clim. Chang.* **6**, 1032–1036 (2016).

7. Global Forest Watch, Global Forest Watch. *https://www.globalforestwatch.org* (2002).

8. F. Schiefer, *et al.*, Mapping forest tree species in high resolution UAV-based RGB-imagery by means of convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **170**, 205–215 (2020).

9. B. G. Weinstein, *et al.*, A remote sensing derived data set of 100 million individual tree crowns for the National Ecological Observatory Network. *Elife* **10** (2021).

10. T. Kattenborn, J. Eichel, F. E. Fassnacht, Convolutional Neural Networks enable efficient, accurate and fine-grained segmentation of plant species and communities from high-resolution UAV imagery. *Sci. Rep.* **9**, 17656 (2019).

11. K. Karra, *et al.*, Global land use / land cover with Sentinel 2 and deep learning in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, (2021), pp. 4704–4707.

12. A. H. Thornhill, *et al.*, Spatial phylogenetics of the native California flora. *BMC Biol.* **15**, 96 (2017).

13. J. P. Smith Jr, California Vascular Plants: A Numerical Conspectus As Seen in Six Standard Floras (2017).

14. B. G. Baldwin, D. H. Goldman, D. J. Keil, R. Patterson, T. J. Rosatti, *The Jepson Manual: Vascular Plants of California* (University of California Press, 2012).

15. B. H. Daru, GreenMaps: a Tool for Addressing the Wallacean Shortfall in the Global Distribution

of Plants. *bioRxiv*, 2020.02.21.960161 (2020).

16. H. Hamilton, *et al.*, Increasing taxonomic diversity and spatial resolution clarifies opportunities for protecting US imperiled species. *Ecol. Appl.* **32**, e2534 (2022).

17. B. H. Daru, *et al.*, Widespread homogenization of plant communities in the Anthropocene. *Nat. Commun.* **12**, 6983 (2021).

18. W. Thuiller, S. Lavorel, M. B. Araújo, M. T. Sykes, I. C. Prentice, Climate change threats to plant diversity in Europe. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 8245–8250 (2005).

19. M. M. Kling, B. D. Mishler, A. H. Thornhill, B. G. Baldwin, D. D. Ackerly, Facets of phylodiversity: evolutionary diversification, divergence and survival as conservation targets. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20170397 (2018).

20. L. J. Pollock, W. Thuiller, W. Jetz, Large conservation gains possible for global biodiversity facets. *Nature* **546**, 141–144 (2017).

21. B. Deneu, *et al.*, Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLoS Comput. Biol.* **17**, e1008856 (2021).

22. T. H. Booth, H. A. Nix, J. R. Busby, M. F. Hutchinson, bioclim : the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. *Divers. Distrib.* **20**, 1–9 (2014).

23. S. J. Phillips, R. P. Anderson, R. E. Schapire, Maximum entropy modeling of species geographic distributions. *Ecol. Modell.* **190**, 231–259 (2006).

24. S. E. Fick, R. J. Hijmans, WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302–4315 (2017).

25. E. Cole, *et al.*, The GeoLifeCLEF 2020 Dataset. *arXiv* (2020) (December 16, 2021).

26. J. N. Pinto-Ledezma, J. Cavender-Bares, Predicting species distributions and community composition using satellite remote sensing predictors. *Sci. Rep.* **11**, 16448 (2021).

27. R. G. Pearson, T. P. Dawson, Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Glob. Ecol. Biogeogr.* **12**, 361–371 (2003).

28. M. S. Wisz, *et al.*, The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological Reviews* **88**, 15–30 (2013).

29. C. F. Randin, *et al.*, Monitoring biodiversity in the Anthropocene using remote sensing in species distribution models. *Remote Sens. Environ.* **239**, 111626 (2020).

30. J. Cavender-Bares, *et al.*, Integrating remote sensing with ecology and evolution to advance biodiversity conservation. *Nat Ecol Evol* (2022) https:/doi.org/10.1038/s41559-022-01702-5.

31. N. Imagery, National Agriculture Imagery Program. *https://www.fsa.usda.gov/programs-and-services/aerial-photography/imagery-programs/naip-imagery/* (December 16, 2021).

32. J. Deng, *et al.*, ImageNet: A large-scale hierarchical image database in *2009 IEEE Conference on*

Computer Vision and Pattern Recognition, (2009), pp. 248–255.

33. T.-Y. Lin, *et al.*, Microsoft COCO: Common Objects in Context. *arXiv [cs.CV]* (2014).

34. B. Deneu, M. Servajean, P. Bonnet, F. Munoz, A. Joly, Participation of LIRMM / Inria to the GeoLifeCLEF 2020 challenge (2020) (February 22, 2022).

35. T. Ridnik, *et al.*, TResNet: High Performance GPU-Dedicated Architecture. *arXiv [cs.CV]* (2020).

36. R. Valavi, G. Guillera-Arroita, J. J. Lahoz-Monfort, J. Elith, Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecological Monographs* **92** (2022).

37. C. R. Keyes, E. K. Teraoka, Structure and Composition of Old-Growth and Unmanaged Second-Growth Riparian Forests at Redwood National Park, USA. *For. Trees Livelihoods* **5**, 256–268 (2014).

38. Golden Gate National Parks Conservancy, Tamalpais Lands Collaborative (One Tam), Aerial Information Systems, Tukman Geospatial LLC, Marin County Fine Scale Vegetation Map (2021) (August 15, 2022).

39. E. N. Stavros, *et al.*, Remote sensing data before and after California Rim and King forest fires, 2010-2015 (2016) https:/doi.org/10.3334/ORNLDAAC/1288.

40. K. Morefield, Ash Creek Wildlife Refuge boundary and ponds, CDFW (2012).

41. K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition. *arXiv [cs.CV]* (2015).

42. K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv [cs.CV]* (2014).

43. E. P. White, *et al.*, Developing an automated iterative near-term forecasting system for an ecological study. *Methods Ecol. Evol.* **10**, 332–344 (2019).

44. GBIF.org (12 May 2022) GBIF Occurrence Download. https://doi.org/10.15468/dl.gt624q.

45. Gadm, GADM database of Global Administrative Areas v 2.0 (2011).

46. A. Paszke, *et al.*, PyTorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32** (2019).

47. C. J. Battey, P. L. Ralph, A. D. Kern, Predicting geographic location from genetic variation with deep neural networks. *Elife* **9** (2020).

48. Szegedy, Vanhoucke, Ioffe, Rethinking the inception architecture for computer vision. *Proc. Estonian Acad. Sci. Biol. Ecol.*

49. Z. Zhang, Improved Adam Optimizer for Deep Neural Networks in *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, (2018), pp. 1–2.

50. A. Guisan, W. Thuiller, Predicting species distribution: offering more than simple habitat models. *Ecol. Lett.* **8**, 993–1009 (2005).

51. B. G. Baldwin, *et al.*, Species richness and endemism in the native flora of California. *Am. J. Bot.* **104**, 487–501 (2017).

52. R. J. Hijmans, S. Phillips, J. Leathwick, J. Elith, dismo: Species distribution modeling. *R package version 0. 8-17* (2013).

Supplementary Materials for:

An image is worth a thousand species: combining neural networks, citizen science, and remote sensing to map biodiversity

Lauren Gillespie[1,2], Megan Ruffley[1], Moises Exposito-Alonso[1,3,4,*]

[1] Department of Plant Biology, Carnegie Institution for Science, Stanford, California, USA
[2] Department of Computer Science, Stanford University, Stanford, California, USA
[3] Department of Biology, Stanford University, Stanford, California, USA
[4] Department of Global Ecology, Carnegie Institution for Science, Stanford, California, USA

[*]to whom correspondence should be addressed: moisesexpositoalonso@gmail.com

Table of contents

An image is worth a thousand species: combining neural networks, citizen science, and remote sensing to track species distributions **1**

1. Building the dataset **5**

Supplemental Table 1 | Key metrics of datasets 5

1.1 Collecting species observations 6

1.1.2 Preliminary dataset 7

1.2 Creating Joint Observations 7

1.3 Generating the test/train splits 8

1.3.1 Interpolation data split 8

1.3.2 Extrapolation data split 8

1.3.3 Train-test split generation for preliminary dataset 9

Supplemental Algorithm 1 | Algorithm for generating the test / train split. 9

1.4 Collecting Remote Sensing Imagery 10

1.5 Bioclimatic Variables 11

1.6 Resolution Limitations of Bioclimatic Variables 11

1.7 Challenges in Using Open-Source Citizen Science Data 11

2. Accuracy Metrics **12**

2.1 Binary classification metrics 13

2.2 Discrimination metrics 14

2.3 Ranking metrics 15

3. Species Distribution Models **16**

3.1 Limitations of on-site sampling methods 16

3.2 Convolutional neural network models 17

3.2.1 Loss functions 17

3.2.2 TResNet CNN architecture 20

3.2.3 Exploring the addition of climate data to CNN models 20

3.2.4 Deepbiospherealpha description 20

3.2.5 Inception V3 Baseline 21

3.3 Climate-only Species Distribution Model baselines 21

3.3.1 Maximum Entropy Baseline 22

3.3.2 Random Forest Baseline 22

3.3.3 Climate-only Multi-Layered Perceptron baseline 23

4. Extracting ecological information from multi-species distribution models **23**

4.1 Validating deepbiosphere's predictions for individual species 23

Table S5 | Site details for individual species case study and human annotation experiments 24

4.1.1 Human annotation protocols and SDM comparisons 24

4.1.2 Redwoods case study 25
4.1.3 Oaks case study 25
4.2 Using deepbiospherealpha to detect spatial ecological processes 25
4.2.1 Detecting local biodiversity 26
4.2.1.1 San Francisco case study 26
Supplemental Algorithm 2 | QGIS query to select for greenspace parks. 27
4.2.2 Detecting community turnover 27
4.2.2.1 Marin case study 28
4.3 Using deepbiospherealpha to detect temporal community changes 29
4.3.1 Detecting successional changes in community composition after fire 30
4.3.2 Detecting deliberate ecological change from community change 31

SUPPLEMENTAL REFERENCES 32

SUPPLEMENTAL FIGURES 37

Fig. S1 | Illustration of joint dataset construction. 37
Fig. S2 | Biases present in the dataset. 38
Fig. S3 | Comparison of different possible remote sensing products. 40
Fig. S4 | Extent of AVIRIS hyperspectral data in continental US. 41
Fig. S5 | Comparison of BioClim and NAIP resolution. 42
Fig. S6 | Location of train and test images used for interpolation validation 43
Fig. S7 | Location of splits used for spatial cross-validation 44
Fig. S8 | Graphical description of image-only TResNet convolutional neural network (CNN) architecture 45
Fig. S9 | Relationship of area under the precision-recall curve with area under the receiver-operator curve per-species 46
Fig. S10 | Comparing area under the receiver-operator characteristic curve by species rank 47
Fig. S11 | Tradeoff between recall and precision per-species across models 48
Fig. S12 | Comparing F1 by species rank 49
Fig. S13 | Example images from human redwood labeling task. 50
Fig. S14 | Aerial imagery used for redwoods human annotation task. 51
Fig. S15 | Correlation of human annotations and deepbiosphere's predictions 52
Fig. S16 | MaxEnt redwoods presence prediction 53
Fig. S17 | Inception V3 baseline redwoods presence prediction 54
Fig. S18 | Deepbiosphere redwood predicted presence at 50m resolution 55
Fig. S19 | deepbiosphere captures interesting co-occurrence patterns in small herbaceous and shrubby plants. 56
Fig. S20 | Predictions of dominant species in the Santa Ynez mountains of Southwest California. 57
Fig. S21 | Comparison of species presence predictions across three species distribution models in Southern California 58

Fig. S22 | Comparison of predicted presence of case study species to CALVEG vegetation mapping. 59
Fig. S23 | San Francisco local biodiversity case study. 61
Fig. S24 | deepbiospherealpha local biodiversity across San Francisco. 62
Fig. S25 | Beta diversity algorithm visual explanation. 63
Fig. S26 | deepbiospherealpha prediction of species turnover (beta diversity) in northern Marin county. 64
Fig. S27 | Relationship of beta diversity to observation density 65
Fig. S28 | Example of change in species presence before and after Rim Fire 66
Fig. S29 | Comparison of deepbiospherealpha community change metric to empirical nDBR metric. 67
Fig. S30 | Decrease in predicted Aspen (Populus tremuloides) presence before and after the Rim Fire. 68
Fig. S31 | Increase in predicted a well-known fire follower perennial herb Blue Dicks (Dichelostemma capitatum) presence before and after the Rim Fire. 69
Fig. S32 | Detecting ecosystem change from wetlands restoration project using deepbiospherealpha. 70
Fig. S33 | Comparing deepbiospherealpha's ecosystem change predictions at Ash Creek National Wildlife Refuge. 71

1. Building the dataset

The methods for dataset collection were inspired from Cole et. al (8) and broadly consist of linking of a set of joint species observations paired with corresponding remote sensing imagery. How the citizen science observations and remote sensing imagery were collected is detailed below.

Supplemental Table 1 | Key metrics of datasets

Main occurrence dataset (*deepbiosphere*)	Observation date range	1/1/2015 - 5/1/2022
	Number of unique species	2,221
	Number of vascular plants	2,221
	Total plant diversity present	29.216%
	Number of unique genera	878
	Number of unique families	153
	Number of unique observations	652,027 images
	Number of linked observations	614,727 images
	Threshold for species inclusion	500 linked observations
	Shannon diversity index	6.825
	Gini inequality index	0.334
	Simpson index	1.000
Preliminary occurrence dataset ($deepbiosphere_{alpha}$)	Observation date range	1/1/2019 - 12/31/2019
	Number of unique species	2,242
	Number of vascular plants	2,195
	Total plant diversity present	28.87%
	Number of unique genera	886
	Number of unique families	181
	Number of unique observations	100,671
	Number of linked observations	89,600
	Threshold for species inclusion	4
	Shannon diversity index	6.7481
	Gini inequality index	0.34272

	Simpson index	0.9999999
Remote Sensing Imagery	Spatial resolution of imagery	1 meter ground sample distance
	Bands used	Blue (428-492 nm) Green (533-587 nm) Red (608-662 nm) Near-Infrared (883-887 nm)
	Year of observation	2012
	Number of unique images	11,095

1.1 Collecting species observations

We collected observations from kingdom *Plantae* using *GBIF.org* from the years 2015-2022 (97). Only records observed by humans with a coordinate uncertainty radius of less than or equal to 120m with no flagged geospatial issues were taken from within the state of California. Nearly all of the subsequent observations were public-derived observations uploaded using the iNaturalist app (9). Any person with a smartphone and who has downloaded the app can upload observations to iNaturalist, meaning that the observations used in this dataset were collected by many thousands of citizen scientists with a wide variety of backgrounds. Accordingly, some observations may be mis-identified.

To minimize mis-identification, only research-grade observations with taxon identifications from at least 2 community members[1] were included in the dataset. That being said, mis-identified observations or mis-located observations can still slip past these filters, making this data especially challenging to work with (11, 12). However, GBIF takes steps to resolve major mis-identification events between closely related taxa (11). Finally, a recent case study from a California island showed that all examined iNaturalist observations with a positional error of < 10m as listed in GBIF were within 270m of the corresponding species detected from remote sensing imagery (12). Extrapolating to this dataset, it's reasonable to assume that the <120m positional uncertainty filter we used would place the vast majority of observations within the linked remote sensing image, therefore preserving the geographic relationship between the observed species and the image.

In total we downloaded a total of 912,380 plant observations of 5,193 unique plant species (97). We further filtered observations to only include vascular plants, which we define vascular plants as all plants in the taxonomic classes of Gnetopsida, Liliopsida, Lycopodiopsida, Magnoliopsida, Pinopsida, Polypodiopsida, Lycopodiopsida, and Ginkgoopsida. We also removed duplicate observations of the same species within a 150m radius, removed species that contain all observations located within a 256m radius, and were not geographically located within the Global Administrative Area boundary of California (40), or were missing climate or NAIP imagery data. To increase the density of observations in the dataset, we used neighbor imputation to add any other species observed within an overlapping 256m radius to a given observation (**Fig. S1**). We finally removed any species that had fewer than 500

[1] https://www.inaturalist.org/posts/39072-research-grade

total observations in the dataset after neighbor imputation, leaving us with a total of 652,027 observations of 2,221 unique plant species (**Table S1**).

1.1.2 Preliminary dataset

Species occurrences were collected from the GeoLifeCLEF dataset (8). These observations were originally downloaded from the Global Biodiversity Information Facility (GBIF), selecting observations within the year 2019, that had a coordinate meter uncertainty of < 30m and fell within the contiguous United States. In total, the dataset contains 100,671 observations for 2,242 unique plant species. The dataset contains both vascular and nonvascular plants, but over 97% of unique species and over 99% of observations are vascular plants (**Table S1**). The original GeoLifeCLEF dataset was further refined to include only observations that fall within the California by assigning each observation to a state using the reverse geocode package and only keeping observations coded within California (13). Furthermore, we added higher-order taxon information to each observation using the GBIF Backbone Taxonomy and only kept observations for species in Kingdom *Plantae* (14). We finally removed any species with less than 5 observations and who only had observations in one geographic location (i.e. within a 256m radius of each other).

1.2 Creating Joint Observations

To create the joint species occurrence dataset, we used neighbor imputation of overlapping observations. Two individual observations from the species observations dataset are considered to be locally overlapping if the Euclidean distance between the latitude and longitude of the two points is less than or equal to 256 m. While this technically means that some neighbor observations may not be geographically located within a neighbor's 256 × 256 pixel image, said resolution is on par with accepted spatial scales in theoretical biogeography and empirical community ecology, which have shown that biotic species interaction networks between individual plants can reach scales of thousands of square meters, and both biotic and shared land use features are thought to drive site-level plant distribution (15–17). Therefore, although there is no strict guarantee that the two observations lie within the extent of their respective NAIP imagery, strong ecological theory supports that the two species may be influencing each other's co-occurrence and subsequent observed local overlap.

The reason that we chose to create a joint dataset is twofold. First, there is theoretical evidence to support that biotic interactions (the interactions of two living species both directly and indirectly) are a strong driving factor in the distribution of species at a variety of scales (16) and that overlap data can be seen as a partial observation of these biotic interactions. For example, Warrior's Plume (*Pedicularis densiflora*) is a hemiparasitic broomrape family member that prefers to parasitize oak and manzanita plants. Subsequently, it is to be expected that many observations of Warrior's Plume should be co-occurring with an oak or manzanita. Second, many species in our dataset have few observations (**Fig. S2A**) which can make it impossible to learn an accurate representation of the species' distribution. However, oftentimes these rarely observed species will inhabit similar habitats to much more commonly observed species, like Warrior's Plume and oak trees. Therefore, providing overlap data may enable the model to bootstrap a better representation for these rare species using shared habitat signatures learned for the more common, overlapping species.

Along with providing overlap data, we also provide higher taxonomic information per-observation. Concretely, the species, genus, and family of all overlapping species is utilized for training. Our

rationale was that the phylogenetic history embedded within the taxonomic hierarchy of species should also encode a shared ecological niche space for some taxonomic groupings.

1.3 Generating the test/train splits

In order to properly validate and compare models, we split the dataset into multiple partitions. Best practices to ensure the reproducibility of machine learning models is to have a train-test-validation split, ideally with the validation data coming from a separate acquisition process to provide as robust a test as possible. To test models with statistical power, validation sets on the order of thousands of observations would be necessary. Furthermore, to test our models appropriately, we require test and validation sets that are at least 1,300m away from all training set examples, to prevent data leakage due to spatial autocorrelation in the coarse resolution bioclim data used by baseline models. Next, since the CNN predicts at 256m resolution, observations with low geographic uncertainty are a must. Finally, since citizen scientist observers using iNaturalist tend to observe species near where they live and on public land they can access, the locations where test observations could be taken from tends to be more remote and inaccessible areas. Needless to say, finding an independent validation dataset that fits all those constraints is a tall order and also changes based on the potential downstream uses of models trained in this way. To that note, to validate the models we used various partitions of the original dataset in a way that tests different properties a well-generalized model should hold, like extrapolation to new geographic areas and interpolation between known observations. We then classify these train/test splits into two types: interpolation and extrapolation, which test the two facets of model performance. The interpolation split of the dataset was used primarily for choosing an optimal loss function and learning rate, while the spatial extrapolations were used to validate these choices.

1.3.1 Interpolation data split

The first partition, which was used for loss function comparison and learning rate selection, was generated by uniformly randomly selecting observations from across the state to be part of the test set and to test interpolation ability (**Fig. S6**). We chose points uniformly across the state to maximize the number of unique environments models would be evaluated on. To ensure the independence of training and testing set data due to spatial autocorrelation, we added all overlapping observations to the test set to guarantee that none of the remote sensing images and observations in the test set were present in the training set. In order to ensure that there was no data leakage between the test and train set only observations which were more than 1,300m away from any other non-overlapping observation were included. We chose an exclusion radius of 1,300m because the climate variable raster pixels converted from arc-seconds to meters can have a diameter of up to 1,200m, so any test set observation within that distance to any observation in the train set would have an identical input value as some observations used during fitting, also known as data leakage. Ultimately 1.88% of the dataset was set aside for testing using this method.

1.3.2 Extrapolation data split

In order to provide cross-validation of the uniform train-test split and to test the extrapolation ability of all models, we also conducted a 5-fold spatial holdout block validation by partitioning California into five 1 degree latitudinal bands (**SM Fig. S7**). Train set points within 1,300m of the test band were removed to prevent data leakage as discussed above. For models utilizing pseudo absence points, all pseudo absence points within the test bands were removed to ensure a fair comparison to

presence-only models. Ultimately, the percentage of test points per-spatial block ranged from 1.40-25.35% of the entire dataset. Since training five models from scratch takes a considerable amount of time and resources, we only compared the final *deepbiosphere* model and all baselines save the Inception baseline CNN model. The Inception model is significantly larger than the TResNet so training takes considerably longer, so we did not include it in this analysis.

1.3.3 Train-test split generation for preliminary dataset

For the preliminary dataset, only one test/train split was generated. To do so, first we calculated the number of observations associated with each species, plus the number of overlapping species associated with those observations. We did so using **Supplemental Algorithm 1**, which ensures that for every observation added to the test set, all overlapping observations are also included into the test set. The algorithm also attempts to balance the train and test such that every species is present in the train set and as many species as possible are represented in the test set. However, due to the strong spatial biases in this type of data, (**Figs. S2**) many observations have few overlapping species while a few observations have many (**Fig. S2A**) and on average species with very few observations often have a lower number of overlapping observations, except for a few rare outliers. Therefore, naively adding observations for these rarely occurring species to the test dataset can often entirely remove other rarely observed species from the training set entirely if observations from said outliers are not chosen carefully.

To combat this, we first add at least one observation for each unique species in the dataset to the test set, then for rarely occurring species which have had all observations moved into the test split, we move all of those observations and overlapping observations back into the train set (**Alg. S1**). We then iteratively add the observation with the fewest number of locally overlapping observations for these rarely occurring species back into the test set, ensuring that an observation can only be added if it doesn't remove another species entirely from the train set.

```
while num species in test split < total number of unique species in
dataset:
    Randomly sample a species with no observations in test set
    Randomly sample an observation for that species
    Add randomly sampled observation to test set
    All all other locally overlapping observations to test split
    # this may move all observations for some species into the test split
For each species with no observations in train split:
    move all of species' observations back to train split
    move all locally overlapping observations back to train split
    # now these species have no observations in the train split
For each species with no observations in test split:
    Find observation with fewest number of overlapping observations
    Add this observation to test split
    Also add all other locally overlapping observations to test set
    If adding observation removes another species from train split:
          Move all observations back into train split
          # these species will not have observations in the test split
```

Supplemental Algorithm 1 | Algorithm for generating the test / train split.

The first stage consists of adding an observation for each unique species. Oftentimes, this will remove all observations of some very rarely observed species completely from the train split. To combat this, we rebalance the split by resampling observations with rare species that also have few

other overlapping observations and species.

1.4 Collecting Remote Sensing Imagery

The remote sensing data used for this project is from the National Agriculture Imagery Project database of aerial imagery from 2012 (18). We chose this dataset because it contains sun angle-corrected orthophotography data collected during the leaf-on growing season with guaranteed < 10% cloud cover at 1 m-resolution (see (19) for a comprehensive overview of NAIP data). We utilized Microsoft's Planetary Computer for dataset access.[2] Briefly, in this dataset the remote sensing images were generated by cropping a 256 × 256 pixel image from the raw NAIP .tiff files centered at the pixel corresponding to the latitude and longitude of each observation (see **Fig S1** for a visual explanation of data collection). Each pixel in the underlying NAIP data is 1m x 1m in resolution (**Table S1**) and each image has four bands corresponding to wavelengths in the red, green, blue, and infrared regions of the light spectrum.

Our decision to work with 4 band RGB-Infrared remote sensing data instead of many-band Landsat data or full-band hyperspectral data was motivated by data accessibility, resolution, and biogeographic theory of scale. While Landsat data is collected worldwide, it has a pixel resolution of 30 m per-pixel, while NAIP data has a 1m-60cm pixel resolution (depending on the year of acquisition), meaning a difference in scale of meters versus kilometers for the images used to train the network (see **Fig. S4** for visual comparison) (19, 20). The resolution of NAIP data compared to Landsat is also more theoretically grounded for modeling species distributions at the meters-level scale. Specifically, recent work in exploring what environmental factors determine the distribution of species at a given spatial scale has shown that at the site (10-1,000 m) scale, land use, soil type, and biotic interactions have dominant control over species distributions, while climate and topography is most dominant at local scales (1-10 km) (17). NAIP imagery certainly contains information about both land use and topography (see **Fig. 1**; **Fig. S1,S3,S5** for examples), and while it lacks explicit soil type information, soil carbon – an important component of soil type – has been successfully predicted using the normalized differential vegetation index (NDVI, an index that can also be calculated from the green and red bands of NAIP) from Landsat data, showing that certain facets of soil type can be captured by high-resolution 4-band remote sensing imagery (21).

Finally, 4-band (RGB-Infrared) satellite imagery is available worldwide at meters-level resolution from various private companies with monthly to daily acquisition cycles, making it easy to scale this framework to continental-level predictions (22). On the other hand, other hyperspectral remote sensing products with a similar level of resolution to NAIP – such as the 224-band, full-spectrum product AVIRIS – have limited coverage at national scale, not even covering all of California (**Fig. S4**). Meanwhile, 4-band remote sensing products of a similar resolution are available from private companies like Maxar Technologies and Planet Labs across the entire globe with weekly to nearly daily acquisitions (22) Therefore, 4-band NAIP imagery provides an optimal mix of data resolution and availability to appropriately model plant communities at a theoretically sound scale with the ability to scale these techniques to the rest of the globe (17).

[2] https://planetarycomputer.microsoft.com/dataset/naip

1.5 Bioclimatic Variables

Beyond remote sensing imagery, we also used WorldClim 2 bioclimatic environmental variables at 30 arc-second resolution (23). We chose to use WorldClim because the 19 bioclimatic variables are a popular choice for bioclimatic SDMs and have been used to generate many hundreds to thousands of SDMs. Using WorldClim as the baseline climatic environmental variables to compare our remote sensing-based approach allows for a fairer comparison to previous methods built using similar features. While methods exist for downscaling these climate variables to a 250m resolution similar to the remote sensing data, all downscaling methods rely on simple interpolation techniques and do not inherently introduce new information about the system (**SM Fig. 4,6**). Therefore, while one theoretically can build bioclimatic variables of a similar resolution to *deepbiosphere's* predictions, fitting an SDM to these higher resolution variables would be synonymous to taking and interpolating an SDM fitted on the 30 arc-second variables. In this work, we choose to use the standard 30 arc-second variables.

1.6 Resolution Limitations of Bioclimatic Variables

While downscaling and interpolation techniques (24, 25) exist that allow one to scale these bioclimatic variables to 250-100 meter resolution – a similar resolution to our methods – these downscaling and interpolation methods are inherently limited in the amount of novel information they can provide, and introduce a new source of potential bias (26). Furthermore, there has been evidence that SDMs trained using coarser-grained climate data overestimate some montane species' tolerance to changing climate compared to meters-resolution local scale information (27). This discrepancy can be easily seen in **Fig. S6** which illustrates the difference between the remotely sensed NAIP imagery and Bioclim at a wetlands restoration site in northeast California. One can easily distinguish the various communities of the wetlands, from the darker gray of the freshwater marsh to the light brown of intruding grasses, along with land use differences between the ecological preserve and the surrounding farmland. This local information is simply not detectable from the projected 12th band of Bioclim data shown. The 30 x 30 arc-second pixels of the bioclimatic rasters are simply too coarse-resolution to capture the local variations between the various communities in this image. Furthermore, as discussed in Section 2.4, from ecological theory it is expected that fundamental drivers of species distribution should be different at local versus site scale, with bioclimatic data driving local scale distribution versus topographic and land use information driving site scale distribution (17). Thus it's to be expected that the information contained in both sources to be radically different.

1.7 Challenges in Using Open-Source Citizen Science Data

Creating remote sensing-based CNN SDMs is challenging due to the large number of species and their unevenness of number of observations per species. Despite our data filtering which included minimum data thresholds for species to be included in the analyses, we still see many dataset imbalances; for example, 20% of species have 1,000 images or less while only 27% of species have 10,000 or more. In addition, 37,300 images contain only one species while only 61,386 images have >100 species, reaching the expected size of a full species checklist for a given 250m radius area. Moreover, species labels are presence-only, meaning that in most cases, the absence of a species label in an image does not guarantee true absence from the ecosystem, only that it has not been observed on iNaturalist at that location. Overall, this leads to a highly imbalanced dataset, with some habitats and some species underrepresented and hard to study.

Specifically, our dataset of observations suffers from three main types of biases. First, there's spatial bias, in that observations are not uniformly distributed across the landmass of California (see **Fig. 1a**, **Fig. S2C,D**). Citizen scientists can only take observations where they can themselves access species, which restricts most observations to publicly owned land and convenience plays a large factor in the distribution of observations, with more observations coming from ecoregions with a higher population or more public parks, such as the Southern California Chaparral or Cascade Range ecoregions (**Fig. S2D**).

Second, observations from casual users tend to show density bias (**Fig. S2B**), where many observed locations have few other overlapping species reported, while a few observations have many overlapping species reported. Oftentimes, this is a result of observers noticing and documenting a particularly salient individual of a species, like when a specific wildflower is in bloom in the spring. However, rarely will users upload all the plant species they may find in a given small area, meaning that the majority of species present at any given observation location in our dataset are unreported. We refer to these present but unobserved species as *pseudo-absences* in our data. The high pseudo-absence rate of our dataset also means that we cannot consider these species occurrences to represent full species checklist data at each site in our dataset, meaning that the co-occurrence network of each site in our dataset is partially observed, again adding extra challenge to the machine learning task.

Third, as mentioned in 1.3 Generating the test/train split, the dataset is also long-tailed in the number of observations per-species, with many species possessing few observations while a few possess a large number of observations in the dataset (**Fig. S2A**). Unfortunately, this "commonness of rarity" is a known phenomenon in plants' distributions (29) and is an expected phenomenon for plant observations. This class imbalance can be problematic for classic machine learning algorithms, since standard accuracy metrics become less informative and models can simultaneously suffer from both under- and over-fitting across classes (30). However, it should be noted that these challenges are not unique to this dataset alone (31–33), thus algorithms successfully able to learn a generalizable representation of this data should be of interest to both the species distribution modeling and more general machine learning fields.

2. Accuracy Metrics

The work in this paper spans many disparate research communities (i.e. deep learning, computer vision, biogeography) which use a variety of different accuracy metrics. To that note, we report eighteen separate accuracy metrics in an effort to be as comprehensive as possible. It should also be noted that what absolute value of an accuracy metric constitutes a "good" model also varies across communities, tasks, and datasets. For example, in the computer vision community many models have been reported to have human-level accuracy on certain tasks, such as image labeling. The take home message is that just because a model does not have a near-perfect accuracy metric does not necessarily mean it's not a good nor useful model. Depending on the use case, even models with low accuracy metrics can still exhibit meaningful and useful behavior, as we demonstrate with the chosen case studies in this work. After all, the definition of a "good model" is derived from the community's consensus on what a good model should be. We hope this work will help contribute to and build consensus in the conversation of what properties and behaviors a good joint species distribution model should exhibit. Due to these limitations within individual metrics, we again emphasize the importance of comparing accuracy across *all* relevant metrics and in this work we consider the "best" models to be those that have competitive accuracy across *all* relevant metrics. For example, a model that has the

highest AUC_{ROCC} but extremely poor binary classification results we believe should be discounted compared to models that are strong –but potentially not globally optimal– across all types of metrics. For the accuracy metrics we group the metrics into three categories to aid in identification and interpretation: binary classification-based metrics, discrimination metrics, and ranking metrics.

2.1 Binary classification metrics

Binary classification metrics are a very common set of metrics used to compare yes/no binary prediction tasks (or e.g. whether a species is or is not present at a given location). For binary classification-based metrics, probability predictions must be converted to a binary presence / absence output using a threshold value. There is vigorous debate within the SDM community on the proper threshold to pick for presence versus absence (51), but for consistency with the computer vision community and the fact that our multi-label domain makes optimal threshold determination non-trivial, we chose to threshold all probabilities ≥0.5 as present and <0.5 as absent. Since our dataset is a multi-label dataset, there are four ways one can partition presence-absences for accuracy calculation across species and across images. We report only per-species metrics in the main text as it provides a richer explanation of the models' ability across species. We calculate these metrics in a multi-label setting, meaning that we use the imputed overlapping species mentioned above in the ground-truth labels when calculating all binary classification-based metrics.

There also exist many different metrics for binary classification but in this work we focus on four common ones: precision, which is a metric of how many species were actually present of the predicted species; recall, which measures how many of the species predicted as present were actually present; and F1, which is the harmonic mean of precision and recall and represents a measure of both skills simultaneously. For *samples* weighted precision and recall, it's possible to calculate the accuracy at each individual image, thus we report both the average and standard deviation across images. For all three *macro* metrics it's possible to calculate the accuracy per-species and per-observation, so we report the average and standard deviation as well. For all metrics, *S* is the number of unique species in the training split of the dataset, N the number of images in the training split of the dataset, $\bar{y}$ is the multi-label ground truth presences and absences of each species for each image, $\hat{y}$ is the SDM's predicted binary present / absent list for each species and each image using a 0.5 threshold, tp is true positives, fn is false negatives and fp is false positives. Per-species metrics were calculated using scikit-learn version 1.1.1 (52) and per-observations metrics were implemented ourselves.

$$\text{per-species recall} = \frac{1}{S}\sum_{i=1}^{S}\frac{tp_i}{tp_i+fn_i}$$

$$\text{per-species precision} = \frac{1}{S}\sum_{i=1}^{S}\frac{tp_i}{tp_i+fp_i}$$

$$\text{per-species F1} = \frac{1}{S}\sum_{i=1}^{S}\frac{2\cdot Rec_i\cdot Prec_i}{Rec_i+Prec_i}$$

For per-observation (i.e. satellite image) accuracy metrics we used the definitions as outlined in Kubany et al (53). We implemented these metrics in Python ourselves.

$$\text{per-observation precision} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\bar{y}_i\cap\hat{y}_i|}{|\hat{y}_i|}$$

$$\text{per-observation recall} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\bar{y}_i\cap\hat{y}_i|}{|\bar{y}_i|}$$

$$\text{per-observation F1} = \frac{1}{N}\sum_{i=1}^{N}\frac{2\cdot Rec_i\cdot Prec_i}{Rec_i+Prec_i}$$

2.2 Discrimination metrics

Binary classification metrics do come with some drawbacks, mainly that the choice of what threshold value to use for determining whether a species is present can have an outsized impact on the accuracy of a model. In order to take into account the effect of thresholds, discrimination metrics measure binary classification ability across a wide range of thresholds in order to both control for the effect of threshold choice and to determine what might be an optimal threshold of presence. Rather than set an arbitrary threshold, they observe how well the model is able to predict across the range of predicted probabilities for that class to see if a higher threshold means more species are correctly predicted without adding in too many false positives. For discrimination metrics, we report the area under the receiver operator characteristic curve (AUC_{ROC}) averaged across species and average area under the precision-recall curve (AUC_{PRC}) (also referred to as mean average precision (mAP) in machine learning papers), along with standard deviation. We again use multi-label, neighbor imputed ground truth presences and absences when calculating discrimination-based metrics. We use scikit-learn version 1.1.1 (52) for all discrimination metrics, which utilizes the trapizoidal integration to calculate area under the curve. $tp(\hat{y}_s, i)$ refers to the number of true positives in $\hat{y}$ when using i as the threshold for predicted presence for species S, while $fp(\hat{y}_s, i)$, and $fn(\hat{y}_s, i)$ are the same for the number of false positives and false negatives, respectively.

$$\text{average } AUC_{ROC} = \frac{1}{S}\sum_{s=1}^{S}\int_{i=\min(\hat{y}_s)}^{\max(\hat{y}_s)}\frac{\left[\frac{tp(\hat{y}_s,i+1)}{\sum_1^N \bar{y}_s} - \frac{tp(\hat{y}_s,i)}{\sum_1^N \bar{y}_s}\right]\cdot\left[\frac{fp(\hat{y}_s,i+1)}{\sum_1^N 1-\bar{y}_s} - \frac{fp(\hat{y}_s,i)}{\sum_1^N 1-\bar{y}_s}\right]}{2}$$

$$\text{average } AUC_{PRC} = \frac{1}{S}\sum_{s=1}^{S}\int_{i=\min(\hat{y}_s)}^{\max(\hat{y}_s)}\frac{\left[\frac{tp(\hat{y}_s,i+1)}{tp(\hat{y}_s,i+1)+fp(\hat{y}_s,i+1)} - \frac{tp(\hat{y}_s,i)}{tp(\hat{y}_s,i)+fp(\hat{y}_s,i)}\right]\cdot\left[\frac{tp(\hat{y}_s,i+1)}{tp(\hat{y}_s,i+1)+fn(\hat{y}_s,i+1)} - \frac{tp(\hat{y}_s,i)}{tp(\hat{y}_s,i)+fn(\hat{y}_s,i)}\right]}{2}$$

One major drawback to this approach is that it does not measure the calibration of an SDM's predicted probabilities well, meaning that a model which has extremely low predicted probabilities can still nevertheless have a high AUC_{ROC} if within its range of predicted probabilities it has good sensitivity and specificity for that class. This means that it's possible to have a model that never predicts a species as present with the standard presence/absence threshold of 0.5 yet still has a high average AUC_{ROC}. Furthermore, tuning the presence/absence threshold using the ROC curve for such models is non-trivial, since the derived optimal threshold will likely be different across species. To correct for this, we also introduce a calibrated area under the curve score where the chosen thresholds are linearly interpolated values between 0 and 1. We use a trapezoidal approximation of area under the curve, utilizing scikit-learn's implementation of area under the curve with 50 uniformly spaced probabilities between 0 and 1.

$$\text{calibrated average } AUC_{ROC} = \frac{1}{S}\sum_{s=1}^{S}\int_{i=0}^{1}\frac{\left[\frac{tp(\hat{y}_s,i+1)}{\sum_1^N \bar{y}_s} - \frac{tp(\hat{y}_s,i)}{\sum_1^N \bar{y}_s}\right]\cdot\left[\frac{fp(\hat{y}_s,i+1)}{\sum_1^N 1-\bar{y}_s} - \frac{fp(\hat{y}_s,i)}{\sum_1^N 1-\bar{y}_s}\right]}{2}$$

$$\text{calibrated average } AUC_{PRC} = \frac{1}{S}\sum_{s=1}^{S}\int_{i=0}^{1}\frac{\left[\frac{tp(\hat{y}_s,i+1)}{tp(\hat{y}_s,i+1)+fp(\hat{y}_s,i+1)} - \frac{tp(\hat{y}_s,i)}{tp(\hat{y}_s,i)+fp(\hat{y}_s,i)}\right]\cdot\left[\frac{tp(\hat{y}_s,i+1)}{tp(\hat{y}_s,i+1)+fn(\hat{y}_s,i+1)} - \frac{tp(\hat{y}_s,i)}{tp(\hat{y}_s,i)+fn(\hat{y}_s,i)}\right]}{2}$$

2.3 Ranking metrics

Compared to binary classifcation and discrimination metrics, ranking metrics focus solely on how high a given species is ranked by probability of presence compared to other species in the same image. These ranking metrics suffer from the same limitations as the aforementioned discrimination metrics in that only compare accuracy of probabilities in a relative sense, rather than the absolute. However, they are staples within the deep learning and computer vision communities, so we choose to report them here for consistency. The first set of ranking-based metrics we report are top K accuracy metrics, a set of single-label metrics. These metrics measure how many times the correct species was correctly predicted within the top K highest-ranked species within a given image. Much like the binary classification metrics, top K accuracy can be calculated across observations and also across species (which we refer to as TopK_{obs} and TopK_{spp}, respectively), and we report both in the supplemental. However, top K accuracy across species is considered to be a better metric of an SDM's ability to distinguish present species, as it corrects for sampling imbalances across species (12, 22).

Machine learning papers oftentimes report top-1 or top-5 accuracy, but given our task is an inherently multi-label one, we choose to report with a larger K than normally seen in computer vision papers as the expected number of unique plant species at the local scale varies anywhere from five to one hundred and thus on average we are most interested in the composition of these top five to one hundred species. To that note, we report both TopK_{obs} and TopK_{spp} for K = 10, 30, and 100. Along with top K accuracy, both top K recall and precision also exist but are far less commonly reported and so we do not report them here. Finally, as a single-label metric, these top K accuracy metrics do not use or capture any information about an SDM's ability to correctly label overlapping species, making them less useful for judging an SDM's ability to capture co-occurrence patterns correctly. We implement this metric in Python using the definitions from (22). Here, $\text{rank}(j, \hat{y}_i)$ is defined as the rank of species j observed in image i from the sorted list of probabilities $\hat{y}_i$ predicted by the SDM, $\text{rank}(j, \hat{y}_i) = \|\{k : \hat{y}_{i,k} \geq \hat{y}_{i,j}\}\|_0$.

$$\text{TopK}_{obs} = \frac{1}{N} \sum_{i=1}^{N} Acc(\hat{y}_{i,j}, K)$$

$$\text{TopK}_{spp} = \frac{1}{S} \sum_{i=1}^{S} Acc_{spp}(\hat{y}_{i,j}, i, K)$$

$$\text{where } Acc_{spp} = \sum_{m=1}^{N_i} Acc(\hat{y}_{m,j}, K) \text{ and where } Acc = \begin{cases} 1 \text{ if } \text{rank}(j, \hat{y}_i) \leq K \\ \\ 0 \text{ otherwise} \end{cases}$$

Finally there does exist a commonly-used multilabel ranking-based metric called label ranking average precision (LRAP) which calculates how highly each species is correctly ranked along with how many other present species are ranked higher, averaged across species. LRAP is the multilabel version of the mean reciprocal rank metric commonly used in document retrieval.

$$\text{LRAP} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\|\bar{y}_i\|_0} \sum_{j:\bar{y}_{i,j}=1}^{J} \frac{\|L(\hat{y}_{i,1:j})\|_0}{\text{rank}(j, \hat{y}_i)} \text{where } L = \begin{cases} 1 \text{ if } \hat{y}_{i,k} \geq \hat{y}_{i,j} \\ 0 \text{ otherise} \end{cases}$$

3. Species Distribution Models

Species distribution models (SDMs) describe how a species is distributed across a given geographic extent. Oftentimes this is accomplished by modeling how the predicted presence of a species across a landscape varies spatially in response to a set of ecologically-meaningful variables. SDMs broadly fall into two rough categories: process-based and correlative. Process-based models attempt to derive fundamental equations for the processes governing a species distribution and build a model of likelihood of occurrence from said processes. Correlative models attempt to infer a species' geographic extent by correlating its known occurrences with a suite of relevant environmental variables and projecting likelihood of occurrence from said correlated variables. However, these definitions are not dichotomous nor mutually exclusive, meaning correlative models may indeed capture some process-based mechanisms intrinsically in their modeling procedure (1).

Further distinction can be made between single species and joint species SDMs, with the former only modeling one species at a time and the latter attempting to model multiple species' distributions simultaneously. Joint SDMs can be further subdivided based upon at what point in the modeling process the species were aggregated. Some joint SDMs are really aggregation of individual SDMs, with the predicted species' presence joined during *post-hoc* analysis, while other joint SDMs model all species simultaneously throughout both the model fitting and analysis steps (2, 3). For future clarification, when referring to a *joint SDM*, we are referring to the latter process.

A final important distinction between different types of SDMs is between functional niche models and realized niche models. The functional niche of a species traditionally was defined as the set of environments where a species individually can sustain itself, while the realized niche is the set of environments where a species can sustain itself in the presence of competition from other biotic sources (4). Contemporary niche theory has strengthened these definitions to include dispersal dynamics, growth rates, and biotic interactions, all vital processes to a species' dispersal (5). However, many modern uses of the terms use a simpler, less precise definition of niche that simplifies the fundamental niche to where a species *can* occur and the realized niche to where a species *does* occur (6, 7). In this work, we aim to build a correlative joint species model that maps where a species do occur in a given 256 × 256m cell.

3.1 Limitations of on-site sampling methods

While the best metric of species presence and current biodiversity comes from on-the-ground observation of species, generating comprehensive checklists of species presence at high resolution across large geographic extents is generally infeasible. For example, in Baldwin et. al. (98), rarefaction curves of species richness were generated at various grid sizes. Using estimates from this analysis of over 1.5 million plant observations, an estimated 1,000 observations would need to be taken per grid scale at a 15 km resolution to reach a point where the rarefaction curve begins to plateau, with many cells requiring upwards of a magnitude more samples due to California's extremely heterogenous and endemic plant communities. Using the 1,000 observation estimations, an conservative estimated 2 million observations would be needed to estimate species richness at 15km spatial blocks, with rapidly increasing numbers of samples as the desired resolution increases. All in all, it would be infeasible to try and perform the same species checklist curation that *deepbiosphere* could theoretically do by hand onsite.

3.2 Convolutional neural network models

Convolutional neural networks (CNNs) are a popular machine learning model that have been adapted to successfully model a wide variety of complex, real-world image-related tasks, from image classification to deciphering handwriting (36, 37). Their widespread success lies in their ability to learn and extract arbitrary patterns and features from images without any human input, allowing them to detect and exploit only the most relevant visual features of an image for a given task. This ability to intrinsically learn the most relevant features of an image makes them very flexible for use across a wide variety of image-related tasks and a wide range of image media, from medical imagery of cells to satellite imagery (38, 39) In this work, we seek to predict the presence of thousands of plant species from relevant features found in the local aerial imagery of a location, a task that CNN models are well-designed to perform. While there have been other deep neural network architectures applied to similar visual tasks, including the U-Net architecture for segmentation and the transformer architecture for video data (38, 40), for the task of *image classification* (the assigning of a discrete label to a given image), at the time of writing the classic CNN architecture still exhibits superior accuracy on ImageNet, a benchmark image classification dataset of similar size and complexity to our own (41). Given that standard classification-based CNNs are more simple, faster to train, and more accessible to a wider array of downstream users than the quadratic memory needs of transformer architectures, we chose to use standard CNN architectures for this work.

3.2.1 Loss functions

The loss, which determines how correct the CNN model's prediction is for a given example, is task-dependent and there are many different choices for a given task type. In this work, we frame our problem as a classification task, where the goal is to classify each image into one of N classes. We also frame our problem as a multilabel task, which extends the above definition to classify each image into K of S classes (where K is the number of present species in the observation). While many CNNs have been developed for image classification, the vast majority of these architectures have been designed for *single label classification*, where for each image exactly one class should apply; for example, each image may either have a dog **or** a cat, but an image is never expected to have both. However, what makes our dataset both unique but challenging is that it provides occurrences of all overlapping species in a given image, making it a *multilabel* dataset (see Section1.2 for more details), since each image is associated with anywhere from one to nearly one hundred overlapping plant species within 256 × 265 m squares. By single-label, we mean that for each training example only the original species observed at that location is included as a positive label, with all imputed neighbors being ignored. When using a multi-label loss, we also include all imputed neighbor species as positive labels in each training example. In all notation $\hat{y}$ refers to the raw outputs from a neural network, $\bar{y}$ refers to the single- or multi-hot vector of known species presence in that example where 0 means a species is absent and 1 means a species is present, and S is the number of unique species in the dataset.

The most commonly used loss function for training classification CNNs in a single label setting is cross-entropy loss (CE).

$$\text{CE} = -L_+ - L_- \begin{cases} L_+ = \bar{y} \cdot log(f(\hat{y})) \\ L_- = (1-\bar{y}) \cdot log(1 - f(\hat{y})) \end{cases} \quad \text{where} f(\hat{y}_i) = \frac{e^{\hat{y}_i}}{\sum_j^S e^{\hat{y}_j}}$$

It should be noted that softmax-based losses like CE loss were designed for classification of single-label datasets like the ubiquitous ImageNet by modeling a probability distribution across labels.

However, this approach does not match the multi-label nature of our task well, as in each image the probabilities across species should be independent so that the presence of one species in an image does not imply a decrease in presence of other species. Furthermore, when using a softmax-based transformation of model outputs, probabilities of an individual species are no longer directly comparable across individual observations on account of the lack of independence across labels, and thus is not a valid transformation for calculating binary classification metrics. Nevertheless, to compare to previous work in (12), we still report accuracy metrics for models trained with CE loss by transforming the model's logits with a sigmoid function rather than a softmax before calculating accuracy metrics which doesn't match the probability distribution CE loss-trained models were trained to match but does fit our task better.

The classic multilabel loss used for training CNNs is binary cross entropy (BCE) loss:

$$\text{BCE} = -L_+ - L_- \begin{cases} L_+ = \bar{y} \cdot log(f(\hat{y})) \\ L_- = (1-\bar{y}) \cdot log(1-f(\hat{y})) \end{cases} \quad \text{where} f(\hat{y}_i) = \frac{1}{1+e^{-\hat{y}_i}}$$

BCE loss can be intuitively interpreted as training the neural network to maximize the conditional log-likelihood of species presence and thus the network's predictions can be interpreted as the likelihood of species occurrence in any arbitrary image. However, our dataset exhibits two important types of imbalance that make this standard loss not ideal for our task. First, our dataset exhibits strong observation imbalance, with few species possessing many observations and many species possessing few observations (**Fig. S2B**) which is problematic since the standard BCE loss formulation assumes an equal number of observations per-class in the dataset. The second is label imbalance, or the imbalance in the number of present vs. absent species per-image, a byproduct of using citizen science observations with incomplete coverage of all species present in a given 256 × 256m area (see Section 2.7 for details). Most observations have fewer than five species present in the observation (**Fig. S3A**), and very few observations reach close to the expected true number of plant species present in a given 512m radius. Therefore, in most cases only one out of one hundred classes from the 2,221 species are present in a given observation and we must assume all locations contain some pseudo absences in $\bar{y}$. Since BCE loss assumes all absences are true absences, yet pseudo absences are guaranteed to be present in the dataset, a loss calculation whose main contributions are from absences points is not ideal for our task.

Therefore, we also considered two other losses that handle the contribution of the negative class differently. The first is a state-of-the-art multilabel classification loss variant of focal loss called asymmetric focal loss (ASL). This loss was explicitly designed and optimized for multilabel tasks by upweighting the loss contributions of the present classes without eliminating the contribution of the absent classes entirely (44). It does so by breaking the standard BCE definition down per-class, depending on whether the class is present or absent in the observation. The contribution of the absent versus present classes is then differentially applied to the loss using the hyperparameters γ_+ and γ_-. By setting $\gamma > 0$, the loss contribution of all absent classes will be scaled down, decreasing their contribution to the overall loss. Furthermore, very easy negative examples which the network assigns low probability ($\hat{y}_i \ll 0.5$) will be exponentially down-weighted, creating a "soft thresholding" effect.

The loss contribution of absent classes can be further reduced for easy negatives through the addition of "hard thresholding" (the function p_m) which fully discards the loss contribution for absent classes with predicted probability below a tunable threshold, m. One can think of this as "throwing away" the loss contribution of very easy classes ($\hat{y}_i \lll 0.5$). Furthermore, the shape of loss function is such that

very hard negative samples ($\bar{y}_i = 0$ when $\hat{y}_i \approx 1$) have a down-weighted loss contribution as well (see Fig. 3 of Ben-Baruch et. al 2020 for specifics). This corresponds to the scenario where an observation is "mis-labeled," which in our dataset would correspond to a location missing an observation of a very clearly present species, thus minimizing the negative effect of density bias seen in our dataset. Finally, the values of γ can be dynamically adjusted during online training to maintain symmetry between the probability contribution of absent versus positive classes, to ensure that the loss contribution of negative samples does not "overwhelm" the contribution of the positive classes. Overall, these benefits fit our dataset well.

This soft and hard thresholding is somewhat analogous to the sampling of pseudo absence points, a required step for the maximum entropy modeling of MaxEnt, but importantly the hyperparameters are not dependent on the spatial extent of a given species, as are the pseudo absence sampling ranges of MaxEnt, and are instead a product of the distribution of presences versus absences in the dataset, thus avoiding the biases introduced through MaxEnt's sampling process (45). We used the recommended default values for $\gamma^+ = 1$, $\gamma^- = 4$ and $m = 0.05$, the default values proposed in (45).

$$\text{ASL} = -L_+ - L_- \begin{cases} L_+ = \bar{y} \cdot (1 - f(\hat{y}))^{\gamma+} \cdot log(f(\hat{y})) \\ L_- = (1 - \bar{y}) \cdot p(f(\hat{y}))^{\gamma-} \cdot log(1 - p(f(\hat{y}))) \end{cases}$$

$$\text{where } f(\hat{y}_i) = \frac{1}{1 + e^{-\hat{y}_i}} \text{ and where } p(\hat{y}) = max(p - m, 0)$$

The second loss function we considered is a novel loss function we refer to as scaled binary cross-entropy loss (scaled BCE) which simply weights the loss contribution of present versus absent classes by the number of present and absent classes, respectively. The effect is that the magnitude of the contribution to the loss of present labels versus absent labels is approximately equal; in other words the correctness of the predictions for present species matters as much as the correctness of the predictions for absent species when calculating the model's regret.

$$\text{Scaled BCE} = -L_+ - L_- \begin{cases} L_+ = \frac{\bar{y} \cdot log(f(\hat{y}))}{\sum_1^S \bar{y}_i} \\ L_- = \frac{(1-\bar{y}) \cdot log(1-f(\hat{y}))}{\sum_1^S 1-\bar{y}_i} \end{cases} \quad \text{where} f(\hat{y}_i) = \frac{1}{1 + e^{-\hat{y}_i}}$$

For completeness, we also compare against ASL loss scaled using the same schema introduced in the scaled BCE loss.

$$\text{Scaled ASL} = -L_+ - L_- \begin{cases} L_+ = \frac{\bar{y} \cdot (1-f(\hat{y}))^{\gamma+} \cdot log(f(\hat{y}))}{\sum_1^S \bar{y}} \\ L_- = \frac{(1-\bar{y}) \cdot p(f(\hat{y}))^{\gamma-} \cdot log(1-p(f(\hat{y})))}{\sum_1^S 1-\bar{y}_i} \end{cases}$$

$$\text{where } f(\hat{y}_i) = \frac{1}{1 + e^{-\hat{y}_i}} \text{ and where } p(\hat{y}) = max(p - m, 0)$$

We compared the performance of each of these losses on the interpolation test set discussed in Section 2.3.1.1 using the image-only TResNet architecture and results can be found in **Table S3**. All losses were implemented in PyTorch and can be found in the code repository associated with the project. We found that our novel Scaled BCE loss had the best performance, highlighting the importance of caring about present versus absent classes equally.

3.2.2 TResNet CNN architecture

Initially we compared a variety of CNN architectures, including VGGNet, ResNet and TResNet architectures (46, 47, 52) and early results showed the TResNet architecture to be superior. This finding is further supported by the TResNet being one of the state-of-the-art architectures for multilabel image recognition tasks. We slightly modified the default TResNet architecture to have four input channels in order to support the infrared band and three fully connected output layers instead of the standard single output layer, corresponding to the three taxonomic ranks being predicted: family, genus, and species (**Fig. S8**, **Table S2**). For learning rate, we tried a variety of static learning rates from 0.01 to 10^{-5} and found a learning rate of 0.0001 to have the highest accuracy on the interpolation test set. However, models trained with ASL loss did slightly better with a slightly higher learning rate, so we report accuracy using a learning rate of 0.0005 for those models.

3.2.3 Exploring the addition of climate data to CNN models

Finally, we were interested in if it was possible to combine both remote sensing and climate sources to improve species modeling using CNNs. Previous work has shown that patterns in climate data can be readily extracted by CNNs (43), however the resolution differences between bioclimatic data and remote sensing data make it impossible to use both types of data simultaneously as separate channel inputs to the CNN, as each 256 × 256 m aerial image is smaller than a single Bioclim pixel (**Fig. S5,** Section1.6). Therefore, we chose to instead construct a novel architecture that combines the extracted features of both the image-only TResNet and the climate-only MLP (Section3.3) during training to navigate the difference in data scales, which we refer to as the joint TResNet architecture (**Fig. S9**, **Table S2**).

The climate-only MLP takes in the raw bioclimatic information pointwise, while the aerial image data is still processed using 2D convolutions. We found that models trained with the climate and aerial image data tended to perform better than models trained with either aerial images or climate variables separately.

3.2.4 *Deepbiosphere*$_{alpha}$ description

While all loss comparison and accuracy metrics reported here are using the dataset with over half a million images, before generating this dataset we used a smaller preliminary dataset extensively with many of the downstream analyses. The CNN used in these preliminary analyses we refer to as *deepbiosphere*$_{alpha}$ which is an image-only TResNet that uses the same architecture found in **Table S2**, except that the output layer is slightly modified to reflect the number of species, genera, and families in the preliminary dataset, which can be found in **Table S1**. This model was trained using the same learning rate and batch size as many of the other SDMs trained using the larger dataset. Since the dataset used to train this model was smaller, the algorithm used to assign observations to the train or test split assigned observations differently from the larger dataset, making direct comparison between the two difficult. In order to compare this model to the others, all observations in the large dataset's interpolation data split that were within 1.3km of the predictions in the training set of deepbiosphere$_{\text{alpha}}$ were not considered when evaluating the model. This left 79% of the observations for evaluation. Furthermore, this model was only trained for 11 epochs due to computational limitations at the time the model was originally trained, plus the size of the dataset used to train the model is less than 1/5th the size of the main dataset. This all leads to this model having considerably poor accuracy on the interpolation data split. However, this does not mean the model is useless,

because it had comparable accuracy to the baseline SDMs during preliminary experiments on the preliminary dataset, and predictions of the redwood and oaks look similar to those from *deepbiosphere*. Rather, we think that the test split used to train this preliminary model was considerably easier since the lower density of observations means that more "easy" observations from locations closer in geographic distance made it into the test dataset. The fact that the exact same architecture trained with the exact same loss function on the main dataset has decent accuracy indicates that the analysis of a preliminary model on the harder, larger main dataset is probably an overly harsh one.

3.2.5 Inception V3 Baseline

Previous work predicting species from NAIP imagery does exist in the form of the GeoLifeCLEF dataset and competition (8). However, this work often employs the standard single-label CE loss that does not match the task of joint species distribution modeling well and focuses on ranking-based metrics like adaptive top 30 accuracy or top 30 error. We argue that this dataset is trying to solve a fundamentally different task from the one we are trying to solve, namely of jointly predicting the presence of multiple species at a time. While the ranking of models trained with ranking-based losses do oftentimes capture some trends in species similarly, the models themselves cannot be readily applied to the downstream tasks we present in this work, as evidenced in **Figs. S17** and **S21**.

The crux of the matter lies with the function used to transform the raw model logits, which can span from $-\infty$ to $+\infty$, to probabilities. One approach uses the sigmoid function to map the domain $[-\infty, +\infty]$ to $[0, 1]$ of each individual value in the prediction. This transformation is therefore independent across values, meaning that an arbitrary number of values can be close to 0 or 1. Conversely, the softmax probability transformation maps each individual prediction to a probability density function, with the restriction that probabilities across classes must sum to 1. This therefore breaks the independence across classes, as if a class is to rise in probability, another class must fall. This means that if a CNN trained with this loss has a high probability that one species is present, for it to also assign probability to another species it's confident is also present, it must decrease the assigned probability to the species already predicted as present. Naturally this makes things messy when wanting to predict upwards of hundreds of species simultaneously. Therefore, this softmax transformation does not fit our task particularly well, as converting probabilities to binary thresholds for calculating accuracy metrics becomes tricky. To this note, we still train and compare against the Inception-based model used in this work, but we only report accuracies for probabilities calculated using the sigmoid transformation. Another point to note is that the original work used a learning rate scheduler. We did not implement a learning rate scheduler in our training framework, so we only report accuracies for the model trained with a constant learning rate of 0.01. We also use the auxiliary loss for model training, which previous work did not, which should improve accuracy by preventing vanishing gradients.

3.3 Climate-only Species Distribution Model baselines

Dozens, if not hundreds, of different SDM methods have been proposed over the decades, ranging from simple linear regression to neural network models (59). Two of the most popular methods, Maximum Entropy (MaxEnt) and random forest (RF), use significantly different methods of prediction and each have been cited thousands of times, warranting their choice as the two baseline SDMs for this study (60, 61). Concretely, we use the popular *dismo* package for species distribution modeling (50) and compare against two popular SDM approaches: Maxent and downsampled single stacked random forest using best practices lined out in (36). We chose these two models specifically

from the dozens of approaches tried in (36) as these two models had consistently the best performance across the hundreds of species in their dataset, besides ensembling approaches. We also attempted to compare against ensembling approaches and run the popular *biomod* ensembling algorithm, but the algorithm was too slow and memory-intensive for us to be able to run it on all species in our dataset.

We used bioclim variables normalized to mean 0 standard deviation 1 as outlined above, and consistent with (36) we removed all but one bioclim variable with a pearson correlation coefficient higher than 0.8, leaving ten variables in total for modeling including Mean Diurnal Range, Max Temperature of Warmest Month, Minimum Temperature of Coldest Month, Annual Precipitation, Precipitation of Wettest Month, Precipitation of Driest Month, Precipitation Seasonality, Precipitation of Wettest Quarter, Precipitation of Warmest Quarter, and Precipitation of Coldest Quarter.

For each species, we generated 50,000 background samples consistent with Valavi et. al. (101) by generating a circular overlay across all points in the training dataset where the radius of each circle is the median distance between observations in the dataset. For the extrapolation experiments using the spatial cross-validation approach outlined above we removed all background samples within the spatially withheld portion of the state. Finally, we used the same background points for both the random forest and Maxent models to ensure a proper comparison.

3.3.1 Maximum Entropy Baseline

For Maxent, we use a stacked single SDM approach to get predictions across the 2,221 species by generating individual models for each species then aggregating the predictions *post-facto*. We use the Maxent implementation from the R package *dismo* using the aforementioned background samples and all presences in a given train split of the main dataset using the 'nothreshold' option. Although studies exist that run MaxEnt on the genus level, we nevertheless opted to only run MaxEnt to model species distribution, not genus nor family. To build the joint prediction across all species for MaxEnt, we then aggregated the predicted score of each individual species from their corresponding projection. Maxent failed to run on 83 species which we exclude when calculating accuracy metrics. Fitting times and memory requirements for the interpolation data split can be found in **Table S3**. For the extrapolation experiments model fitting on all 11 splits of the dataset was done using 24 cpus and 128 GiB of RAM per-split for a wall clock time of around 8 hours to fit maxent. However, without parallelism, fitting maxent would take almost two days to fit and without parallelism across the dataset splits, would take an estimated over 20 days to fit maxent.

3.3.2 Random Forest Baseline

For the random forest baseline, we also take a stacked single SDM approach and use *dismo*. For each species we fit a random forest with 1,000 trees using equal bootstrapping of positive and negative samples with replacement. All other options were the default settings that come in *dismo*. 66 species did not properly fit for random forest, and those were excluded from subsequent analyses. Fitting times and memory requirements for the interpolation data split can be found in **Table S3**. For the extrapolation experiments model fitting on all 11 splits of the dataset was significantly faster with random forest and done using 24 cpus and 128 GiB of RAM. It took about an hour to fit random forest with parallelism and an estimated day and a half to fit all models sequentially.

3.3.3 Climate-only Multi-Layered Perceptron baseline

To compare the difference in remote sensing data versus standard bioclimatic data for species distribution modeling, we also considered how well a standard fully-connected multi-layered perceptron (MLP) SDM would perform trained on Bioclim. MLPs trained on environmental data are a standard choice for SDMs, although Maxent and random forest have been more popular in recent years (58). To do so, we implemented a two layer, fully-connected MLP with BatchNorm (**Table S2**). Specifically, the model's architecture was inspired by (42) and consisted of two fully-connected layers with 1,000 neurons each, followed by a dropout layer with a 0.25 dropout rate, then by two layers with 2,000 neurons each, before predicting species, genus, and family (**Table S2**). Batch size, total number of epochs, memory usage and training time can be found in **Table S3**. It should be noted that this MLP is *not* convolutional and learns from the raw value of environmental variables, rather than two-dimensional patterns of remote sensing data, like the CNN models do. The climate-only MLP was also trained using the Scaled BCE loss function discussed in Section3.1.1.

4. Extracting ecological information from multi-species distribution models

While the ability to successfully predict the presence of thousands of species at once is in itself an impressive feat, also of importance is *how* such an SDM can be useful for a variety of downstream ecological tasks. By modeling each individual species of a community, we can begin to detect not just individual species, but phenomena of the entire community as well. While traditional SDMs have been used to model entire genera simultaneously, or single species SDMs have been aggregated across thousands of species to create continent-wide metrics (2, 24), and while CNNs have been used to detect vegetation categories and Ellenberg's indicator values (65, 66) using a single SDM to detect community-level patterns from individual species distribution data has not been attempted to date. Here, we explore how both individual species-level phenomena are learned by *deepbiosphere* and how community ecological level phenomena and change are also well-captured by *deepbiosphere*.

4.1 Validating deepbiosphere's predictions for individual species

Since *deepbiosphere* is a joint SDM, it can be used to detect and model individual species at a time, along with the broader community. However, how to validate the accuracy of the model for a given species' range is difficult without on the ground, gold-label observations. Barring onsite validation, we opted to instead compare how well *deepbiosphere* could classify species presence for two large charismatic tree species –Redwoods and Valley Oak– compared to humans attempting the same task. Specifically, we compared *deepbiosphere's* predictions to humans labeling the exact same aerial imagery that *deepbiosphere* predicted from. We specifically chose these two species as they are observable directly from the NAIP imagery so that human annotators could distinguish their specific canopies and the human annotations could be reasonably assumed to represent some proxy of ground-truth presence and absence, plus these species have a decent number of observations in the dataset, making it more likely *deepbiosphere* learned a good representation of these species' distributions.

We picked two locations within both species' predicted range on Calscape[3] as case studies. For the redwoods case study, we chose the location based on known remaining old-growth redwood groves (old-growth groves harbor the largest redwoods, making them most easy to detect from remote sensing imagery) and chose Tall Trees Grove in Redwoods National and State Parks as the case study location (**Table S3**). For the oaks case study, we searched for *Quercus lobata* occurrence records in Calflora, a plant occurrence database independent from ours, and selected a site with multiple observed oaks in an undersampled region in our dataset (**Table S5**) (67). For each site, we selected the most centered NAIP imagery tile as the extent to compare predictions. Each tile is approximately 5 x 6 km in extent and an example can be seen in **Fig. 2A**.

Table S5 | Site details for individual species case study and human annotation experiments

Redwoods case study location: (41.209, -124.01)		**Oaks case study location:** (34.533, -120.17)		
Redwoods example locations	Grove – Park	Oaks example locations	Species	Calflora ID
(40.3527, -123.9894)	Bull Creek Flats – Humboldt Redwoods	(36.108322, -120.561226)	*Quercus lobata*	po68973
(41.7564, -124.1087)	Grove of Titans – Jedediah Smith Redwoods State Park	(36.151438, -120.770939)	*Quercus lobata*	po68984
(40.6554, -124.0998)	Elk River Trail Grove – Headwaters Preserve	(36.635690, -121.242532)	*Quercus lobata*	po122185
		(34.538051, -120.191470)	*Quercus agrifolia*	po60499
		(34.540338, -120.158243)	*Quercus agrifolia*	po63343

4.1.1 Human annotation protocols and SDM comparisons

In order to generate ground-truth human annotations to compare *deepbiosphere's* predictions to, we generated a user study implemented in Google Sheets where human annotators classified the same aerial imagery as *deepbiosphere* by percent cover. Annotators were not domain experts and the only training received *a priori* were three already classified example images. These already classified images were taken from the same 2012 NAIP acquisition to ensure standard data quality. These examples were deliberately chosen to be outside of the case study area but within the species' core ranges and were annotated on a scale of 0% cover to 100% cover, with 5 possible categories (**Figs. S14, S20B**). The locations for each example image can be found in **Table S5**, along with the source of the example (either known old-growth grove maps or Calflora) (68). The locations for oaks were

[3] https://calscape.org/

verified *Quercus lobata* observations taken from Calflora (**Table S5**). In total, three annotators annotated the redwoods case study and two annotators annotated the oaks case study.

Annotators were then given the NAIP imagery partitioned into 256 × 256m blocks (**Fig. S13**), which were labeled A-Z and 1-30 to correspond with the appropriate cell in Google Sheets. Three annotators annotated the redwoods example and two annotated the oaks example. There was some small discrepancy between annotator's assessments so we took the average annotated value for comparison. CNN models from the last epoch of training for *deepbiosphere* and the Inception baseline were used, rather than the epoch 19 models used to evaluate performance (epochs were standardized and some models were only trained 20 epochs in total).

4.1.2 Redwoods case study

For the redwoods case study, human annotations were compared to *deepbiosphere* predictions (**Figs. 2D, S15**) MaxEnt baseline predictions (**Figs. S16**), and Inception baseline predictions (**Figs. S17**) by correlating the two values and determining significance using the modified t-test from SpatialPack to correct for spatial autocorrelation between nearby pixels (69). Cohen's kappa and Spearman's rank correlation are also reported for some models using scipy implementations. Locations for spatial autocorrelation correction were determined by taking the centroid of each pixel. Co-occurring species *Oxalis oregana* and *Vaccinium parvifolium* were chosen based on previous work comparing old-growth and secondary-growth redwoods (96).

4.1.3 Oaks case study

For the oaks case study, alliance-level vegetation classification for the extent of the study region was downloaded from the USDA's FSgeodata clearinghouse portal (99). Descriptions of vegetation alliances were taken from the U.S. Forest Service's CALVEG zone description manual (100). Species were chosen based on if there was an alliance in the case study area that they were the dominant species of and if a substantial area was represented, excluding valley oaks. The chosen species were valley oak (*Quercus lobata*) associated with the valley oak alliance (QL), coast live oak (*Quercus agrifolia*) associated with the coast live oak alliance, and buckbrush (*Ceanothus cuneatus*) associated with the Ceanothus chaparral alliance. Each cell of the image was then labeled as either within or outside of each alliance using geopanda's sjoin algorithm between the vegetation map and model predictions. For the *deepbiosphere* and *Inception* predictions, predictions were made at 50m resolution to better capture the quick transitions between habitats that don't resolve at 256m resolution. Predictions were compared per-SDM and per-species using an unpaired student t-test on whether the shape was annotated as within or outside the associated vegetation alliance. Boxplot width is mapped to the number of shapes for that treatment, which is higher for the higher-resolution predictions. Human predictions were also included to see what species they were associated with.

4.2 Using $deepbiosphere_{alpha}$ to detect spatial ecological processes

On top of modeling individual species, *deepbiosphere's* joint species predictions can also be thought of as modeling entire plant communities simultaneously by capturing patterns in communities' individual

constituents. The result is that community-level ecological metrics can be observed from *deepbiosphere's* predictions, enabling the direct detection of important ecological processes such as biodiversity and community turnover. Historically, these processes have been difficult to observe directly without detailed species checklist data, limiting the resolution at which these metrics can be generated (70, 71). However, combined SDMs for thousands of species have been proposed previously to measure certain important ecological phenomena, such as imperiled species index (2). Here, we use *deepbiosphere's* predicted presence for plant species to generate maps of local biodiversity and community turnover. Specifically, we use the preliminary model $\text{deepbiosphere}_{\text{alpha}}$ for these analyses.

4.2.1 Detecting local biodiversity

In order to interpret *deepbiosphere's* predictions of species presence as local biodiversity, the CNN's predicted probabilities must be thresholded to turn said probabilities into binary presence / absence maps. How predicted SDM presences should be thresholded is an active area of discussion and research in the ecological community (75, 76), but in the machine learning community, thresholding probability values for binary classification is standard at $\hat{y} \geq 0.5$ = present (35). Therefore, for simplicity and consistency we thresholded $\textit{deepbiosphere}_{alpha}$*'s* predicted probabilities with the standard 0.5 threshold. To calculate alpha diversity (refering to species richness within a quadrant, and not to be confused with $\textit{deepbiosphere}_{alpha}$), we then count the number of species above this threshold for a given image, namely:

$$\text{alpha diversity} = \sum_{i=1}^{S} \mathbf{1}[\hat{y}_i \geq 0.5]$$

Validating alpha diversity is difficult without full species checklist data which we could not find at the scale necessary to achieve statistical significance. Therefore we once again present a case study and employ proxies to showcase the potential of this method. Specifically, we focus on an urban area to highlight a potential use case of these predictions for downstream decision making tasks.

4.2.1.1 San Francisco case study

We chose to focus on the San Francisco peninsula for our case study, as it is an urban area that nevertheless possesses many naturalized areas nearby. Intuitively, industrialized and highly developed areas will have less natural biodiversity, being highly disturbed and with little bare soil for plant growth. However, San Francisco also possesses many nearby and urban parks, some of which are recovering landscapes and others which are more pristine habitats, both of which are expected to have a higher biodiversity than a city block. This close proxy of highly urbanized environments near native habitats makes it easy to visually contrast the predicted alpha biodiversity and to detect localized areas of high biodiversity and low biodiversity all within a small area.

We specifically focus our analysis on the counties of San Francisco, south Marin, and extreme northern San Mateo (**Fig. S23A**) as they form the urban core of San Francisco. The case study region extent is around ~10 x 40 km, and is constrained by the available NAIP imagery, which is why the extreme western and eastern ends of the peninsula are cropped out. To confirm our intuition that city blocks should have a lower biodiversity than green zones in urban areas, we classified each 256 × 256m block as either an urban area or a green space, depending on if the area falls within a park or not. To do so, we acquired the boundaries of all public parks in San Francisco, San Mateo, and Marin counties (**Fig. S23B**).

Specifically, for Marin county, we utilized publicly available maps of the county's digitized parks and greenbelts (77, 78). After merging the two maps in QGIS 3.8, since the map for Marin's parks included schools, industrial parks, and other non-green space parks, we filtered the dataset down to only natural area parks using QGIS' query builder with the filtering formula from **Suppl. Alg. 2** (79). We also used QGIS' "fix geometries" option to resolve geometric inconsistencies.

```
"ParkType" != 'City Park' AND "ParkType" != 'Neiborhood park_' AND "ParkType" ILIKE '%park%' OR "ParkType" ILIKE '%open%' OR "ParkType" ILIKE '%watershed%'
```

Supplemental Algorithm 2 | QGIS query to select for greenspace parks.

For San Francisco county, we utilized a public map of land use categorization for each parcel in the county, specifically keeping parcels coded as "open space," and a public map of all state park boundaries in California (80, 81). However, parks and open spaces owned by the National Park system, the San Francisco Port Authority, and the University of California system were not present in the original set of open space parcels. To account for this, we also added hand-drawn boundaries for the Mount Sutro Open Space Reserve, Warm Water Cove park, Heron's Head park, Alcatraz island, and Pier 94 wetlands as presented in OpenStreetMaps (82). No filtering of the parks in San Mateo county was necessary.

Once each individual map layer had been pre-processed, we merged each of the aforementioned features using QGIS's built-in "merge vector layers" and "dissolve" functions. Finally, we clipped all the merged features using the Database of Global Administrative Areas level 1 boundary map of California and the outline of the extent of the case study area (83). To compare deepbiosphere's alpha diversity predictions to the open spaces map discussed earlier, we vectorized the alpha diversity predictions for each 256 × 256m square in the study extent, then used rasaterio's "mask" algorithm to determine if any features from the parks map intersected with each pixel's boundary. A pixel was marked as belonging to a green space if the center of the pixel resided in one of the features from the parks map, else it was marked as an urban area. The pixel values were then compared using an unpaired student's t test.

Since iNaturalist users can only upload species observations in locations they can physically access, we also compared how strong the relationship between observation density and alpha diversity was. Using both train and test observations from our dataset (**Fig. S24B**), we counted the number of observations that were located within an alpha diversity pixel using Geopanda's "intersect" algorithm (84). We then correlated the number of observations with the alpha diversity using a standard Pearson's correlation implemented in the core R stats library (**Fig. S24C**).

4.2.2 Detecting community turnover

In order to interpret *deepbiosphere's* predictions of species presence as community turnover, we utilized a novel edge detection algorithm inspired by common edge detection filters from computer vision, such as the Sobel and Laplacian filters (**Fig. S25**). At a high level, since each pixel of the predicted presence map *deepbiosphere* can be interpreted as the alpha diversity (see Section 4.2.1 for details), we can then interpret the *change* across neighboring pixels as an approximation of the difference between the two pixels, and thus interpret it as the beta diversity. To avoid problems with picking threshold values to determine whether a species is present or absent in a given pixel (see 4.2.1

for details of this discussion) we use distance in probability space as the proxy for change. Specifically, we take the Euclidean norm of the difference between the predicted probability vectors for two neighboring pixels, $\hat{y_1}$ and $\hat{y_2}$ as a proxy for community turnover (**Fig. S25B**).

$$\text{Euclidean norm} = \sqrt{\sum_{i=1}^{S} (\hat{y_{1_i}} - \hat{y_{2_i}})^2}$$

We chose the Euclidean norm as it is a metric that encodes both magnitude and direction between the two vectors. Therefore, it will capture both changes in a given species' presence across pixels through the directional component (e.g: a species highly predicted in one pixel has much lower predicted probability in the other pixel) and capture changes in the raw number of species predicted through the magnitude component (e.g.: many species are predicted with high probability in one pixel, but then very few are predicted with high probability in the other pixel).

Generalizing, for all non-edge pixels in the given extent of *deepbiosphere's* presence predictions (**Fig. S25A**), we take the Euclidean norm from the central pixel to its neighboring pixels, then take the average across its eight neighbors to generate the final beta diversity value (**Fig. S25B-D**). This algorithm is convolutional in the sense that the same operation is performed by "sliding" the neighbor window down and across pixel-by-pixel. Finally, pixels with a higher intensity of this beta diversity metric can be thought of as detecting places of ecosystem transition, as a higher beta diversity means more rapid or intense change within the local community writ large.

4.2.2.1 Marin case study

Much like alpha diversity, finding adequate data to directly validate *deepbiosphere's* predictions of beta diversity and community turnover is exceedingly difficult. We again examine a case study in north Marin county of California and use proxy data to explore how well *deepbiosphere* captures facets of species turnover (**Fig. S26B**). We chose this region as it has a decent number of observations, a highly varied landscape, plus it sits at the transition point between two ecoregions of California, and is thus a known site of ecosystem transition (**Fig. S26A**). Another boon is that in the area, the vegetation was classified down to the association level at a meter level resolution (**Fig. S26C**), giving us fine-grained access to community-level classifications of the present ecosystems (86). Again, the actual extent of the case study area was determined by the available NAIP imagery sections, leading to a 15 x 18km region for analysis.

In order to compare *deepbiosphere's* beta diversity predictions to a proxy of ground-truth community change, we compared the beta diversity predictions to the number of unique vegetation classes present in each pixel. To do so, first we fixed all invalid polygons in the fine scaled vegetation classification map using Shapely's "make_valid" function (87). Next, we cropped the vegetation map features to the extent of the area, then merged all polygons of the same vegetation classification using Geopandas' "dissolve" function (84). For each vectorized predicted beta diversity pixel, we then calculated how many of the merged unique vegetation associations overlapped that pixel's boundaries using Geopanda's "overlaps" function (84). We then correlated the number of overlapping vegetation classes with the beta diversity per-pixel, correcting for spatial autocorrelation using the modified t test from SpatialPack, again using the centroid of each pixel as the coordinates per-sample (**Fig. S26D**) (69).

Similar to the alpha diversity analysis, we were curious to see if observation density was potentially a driving factor in *deepbiosphere's* beta diversity predictions. To explore this question, we took all

occurrences from within the case study area (**Fig. S27**) and correlated it with the predicted beta diversity from *deepbiosphere* (**Fig. S27B**). We found no relationship between the two variables, indicating that the predicted beta diversity is not a degenerate factor of observation density in the area.

4.3 Using *deepbiosphere*$_{alpha}$ to detect temporal community changes

One major benefit to using remote sensing data to train SDMs as compared to climatological datasets like WorldClim is that remote sensing data often has a shorter time period between acquisitions. For example, WorldClim provides bioclimatic variables for one fixed time point for the current historical period, along with future averages in 20 year intervals (23). Capturing temporal ecological changes due to natural events such as drought, fire, or flooding is simply not possible at such a coarse temporal resolution. Meanwhile, NAIP data is acquired every two years, while a full Landsat acquisition of the earth happens roughly every 16 days (18, 20). With this fine-grained temporal resolution, ecological changes year-to-year and even season-to-season are potentially detectable. Here, we explore how predicted species presence by *deepbiosphere* changes between 2012 and 2014 in a case study highlighting successional community shifts and another highlighting deliberate ecological change. We chose these two examples to highlight how *deepbiosphere* can capture temporal change within and across ecosystems, an ability that will be paramount to help monitor our changing ecosystems under the Anthropocene.

In order to detect temporal change, we employed a similar method to the beta diversity algorithm discussed in Section 4.2.2. Specifically, the metric of community change we used was the Euclidean distance between *deepbiosphere's* predicted species probabilities in 2012 and 2014.

$$\text{Community change} = \sqrt{\sum_{i=1}^{S} (\hat{y}_{i\ 2012} - \hat{y}_{i\ 2014})^2}$$

We chose Euclidean distance as the metric of choice for similar reasons as discussed in Section 4.2.2, mainly that it captures both the magnitude and direction of change. Other metrics considered was the change in alpha diversity across time as defined in Section 4.2.1, which we declined to use since the choice of threshold value has a significant and potentially arbitrary impact on the number of present species. We also considered the sum of probabilities across classes, which would be the same calculation as alpha diversity minus the binary thresholding, but this metric only captures the magnitude of change and fails to account for direction, thus our decision to ultimately settle on Euclidean distance.

We chose the years 2012 and 2014 as *deepbiosphere* was trained on imagery from 2012 and because the 2014 NAIP imagery was the last year imagery was acquired at 1m resolution. However, this analysis can be compared across any range of years which remote sensing imagery of the right resolution, either through acquisition at the same resolution or upsampling to the same resolution as the model was trained with. One could also theoretically predict with images of higher resolution using 256 × 256 pixel windows at a time, however this dataset shift would likely lead to worse and/or incorrect predictions as the size of topographical features in the higher or lower resolution imagery would no longer match the patterns learned by the filter banks within the model. The same dataset shift phenomenon is also potentially a factor for predicting with imagery from a different year even if it is the same resolution, as differences in sensors used to collect the imagery can lead to a higher or lower mean pixel value. However, mean centering the data helps to reduce this sensor shift.

4.3.1 Detecting successional changes in community composition after fire

Fire is a natural part of many of the ecosystems of California, with a series of unique plant communities that succeed one another following a fire event. However, almost a century of fire suppression has disrupted the cyclic nature of these successional changes in community composition in ecosystems adapted for fire, altering these different steps in the succession and even leading to more intense fires in California (88, 89). We were curious to see if *deepbiosphere*$_{alpha}$ could detect the expected community composition change that comes after a fire.

To this point, we focused in on the Rim Fire of 2013 as our case study, as at the time of the blaze was California's second-largest fire on record, occurred in the intervening years between NAIP acquisitions, and occurred in a part of California with few observations, with the dataset *deepbiosphere*$_{alpha}$ was trained on only possessing 81 training observations inside the fire perimeter. More specifically, the blaze occurred in the Hetch Hetchy Valley in the western Sierra mountains, and happened in the fall of 2013, giving an entire growing season between the fire and the acquisition of the next round of NAIP imagery in summer of 2014. We used the community change metric outlined above, with an example of the underlying data for this algorithm found in **Fig. S28**. In order to reach the same resolution as the empirical burn severity data outlined below, we convolved *deepbiosphere*$_{alpha}$ across the NAIP imagery by sliding the predictions across eight pixels at a time to produce predictions at 35m resolution.

In order to compare deepbiosphere's predictions to some empirical approximation of community change, we acquired a dataset of difference in normalized burn ratio (dNBR) observations from Stavros et. al. (90). Normalized burn ratio (NBR) is an empirical measurement of burn severity calculated by taking the difference between a near infrared wavelength – which captures photosynthesis intensity – and a shortwave infrared wavelength –which captures heat absorption from char – which is typically measured using hyperspectral spectrometers such as AVIRIS or MASTER sensors (91). To get dNBR, the change in NBR from data acquired before the fire to data acquired after is taken.

For this analysis, we chose to use MASTER sensor data as it had more coverage over the fire than the AVIRIS sensor data, covering roughly half of the fire's extent. We also chose to look at dNBR calculated using NBR acquired in June of 2014, as that was the closest time point to the acquisition of the corresponding NAIP imagery. We also used the provided perimeter shapefile for determining the perimeter of the fire in all subsequent analyses. Both deepbiosphere's community change prediction and the empirical dNBR can be seen side-by-side in **Fig. S29**.

To compare deepbiosphere's predicted community change metric to dNBR, we performed a Pearson's correlation, correcting for spatial autocorrelation using the Dutilleul correction from the SpatialPack R package (69, 92). Before correlation, we averaged both *deepbiosphere*$_{alpha}$*'s* predictions and the dNBR data across 8 x 8 block chunks to produce values at 280m resolution, close to the default resolution of *deepbiosphere*$_{alpha}$*'s* predictions. We averaged the values because at such a high resolution the pixels may not line up perfectly with one another due to small differences in how the data is overlapped. However, the R^2 does not change substantially with averaging versus pixel-wise comparison (an R of 0.47 compared to 0.54, a difference of approximately 0.07) showing that correlation is not strongly dependent on scale.

Finally, we also compared the change in predicted probabilities for two charismatic species –Aspen and Blue Dicks– to show how *deepbiosphere*$_{alpha}$ appears to capture expected changes in abundance

before versus after the fire (**Figs. S30, S31**). To do so, we compared the signed difference in probability for each species' predicted presence before versus after the fire. We chose this to highlight how a species' predicted presence increases or decreases after the fire.

4.3.2 Detecting deliberate ecological change from community change

Unlike the successional community changes that characterize the cyclic nature of fire-prone ecosystems in California, some community changes are permanent, reflecting a change in the underlying ecosystem of a given area. Sometimes these changes are deliberate, such as land use changes for agriculture or for habitat restoration, other times they are not, such as desertification. Regardless of the cause of change, an ecosystem undergoing a permanent or semi-permanent change should naturally see this change reflected in its community composition over time, as organisms adapted to this new environment outcompete and replace the existing community members. Therefore, we might expect that some areas of increased community change as predicted by $deepbiosphere_{alpha}$ should reflect a marked ecological change undergoing in that area. However, oftentimes ecosystem change is a relatively slow process yet some deliberate changes can occur on the timescale of seasons to years.

In order to see if increases in predicted community change could reflect an underlying deliberate ecosystem change, we searched for examples of habitat restoration in California between 2012 and 2014. In the northeast part of the state, we found a state wildlife area on the border of Modoc and Lassen county called Ash Creek Wildlife Area which protects a rare seasonal freshwater wetland. Said wetland had been degraded by nearby agricultural activities and farmers would cut and bank channels in the natural marshland banks to prevent their fields from flooding during the growing season. The impact of these cut channels, coupled with recent strong droughts has led the water table to decrease and sharply restricted the area in which restorative spring rainfall can pond and create the tule marsh, allowing non-native annual grasses to begin to encroach into the wetlands (93).

In order to restore the imperiled wetlands, the preserve began undertaking a series of ecological restorations starting in 2013, which involved a "pond and plug" strategy (93, 94). This strategy involves digging ponds in the grassland areas and use the excavated dirt to shallow out and flatten the cut channels, allowing restorative water flow back into the marsh during winter rains. During the initial 2013 phase, 80 ponds were added and 1,200 acres of wetland restored after the following winter rains.

We explored deepbiosphere's ability to detect this deliberate ecosystem change by looking at the predicted community change between 2012 and 2014, as outlined above (**Fig. S32**). The outline of the preserve was acquired from the California Department of Fish and Wildlife (95) and pond outlines were acquired from California Department of Fish and Wildlife (94). We also considered if the change in predicted community change increased significantly inside the preserve versus outside, and within the ponds versus outside using an unpaired student t-test to compare the per-pixel values of community change inside versus outside the park and pond boundaries (**Figs. S33**).

SUPPLEMENTAL REFERENCES

1. C. F. Dormann, *et al.*, Correlation and process in species distribution models: Bridging a dichotomy. *J. Biogeogr.* **39**, 2119–2131 (2012).
2. H. Hamilton, *et al.*, Increasing taxonomic diversity and spatial resolution clarifies opportunities for protecting US imperiled species. *Ecol. Appl.* **n/a**, e2534.
3. M. W. Tobler, *et al.*, Joint species distribution models with species correlations and imperfect detection. *Ecology* **100**, e02754 (2019).
4. J. Roughgarden, The Fundamental and Realized Niche of a Solitary Population. *Am. Nat.* **108**, 232–235 (1974).
5. W. Godsoe, J. Jankowski, R. D. Holt, D. Gravel, Integrating Biogeography with Contemporary Niche Theory. *Trends Ecol. Evol.* **32**, 488–499 (2017).
6. K. C. Rosenblad, D. L. Perret, D. F. Sax, Niche syndromes reveal climate-driven extinction threat to island endemic conifers. *Nat. Clim. Change* **9**, 627–631 (2019).
7. E. A. Beever, *et al.*, Improving Conservation Outcomes with a New Paradigm for Understanding Species' Fundamental and Realized Adaptive Capacity: A new paradigm for defining adaptive capacity. *Conserv. Lett.* **9**, 131–137 (2016).
8. E. Cole, *et al.*, The GeoLifeCLEF 2020 Dataset (2020).
9. C. A. Matheson, iNaturalist. *Ref. Rev.* **28**, 36–38 (2014).
10. J. P. S. Jr, California Vascular Plants: A Numerical Conspectus As Seen in Six Standard Floras. 3.
11. G. Samy, *et al.*, Content assessment of the primary biodiversity data published through GBIF network: Status, challenges and potentials. *Biodivers. Inform.* **8** (2013).
12. K. A. Uyeda, D. A. Stow, C. H. Richart, Assessment of volunteered geographic information for vegetation mapping. *Environ. Monit. Assess.* **192**, 554 (2020).
13. A. Thampi, Reverse Geocoder (2022) (March 16, 2022).
14. , GBIF Backbone Taxonomy https:/doi.org/10.15468/39omei (March 16, 2022).
15. Y. L. Dupont, K. Trøjelsgaard, J. M. Olesen, Scaling down from species to individuals: a flower–visitation network between individual honeybees and thistle plants. *Oikos* **120**, 170–177 (2011).
16. M. S. Wisz, *et al.*, The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biol. Rev.* **88**, 15–30 (2013).
17. R. G. Pearson, T. P. Dawson, Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Glob. Ecol. Biogeogr.* **12**, 361–371 (2003).
18. Aerial Photography Field Office, National Agriculture Imagery Program (NAIP) (2012) (May 26, 2021).
19. A. E. Maxwell, T. A. Warner, B. C. Vanderbilt, C. A. Ramezan, Land Cover Classification and Feature Extraction from National Agriculture Imagery Program (NAIP) Orthoimagery: A Review. *Photogramm. Eng. Remote Sens.* **83**, 737–747 (2017).
20. US Geological Survey, National Aeronautics and Space Administration, Landsat 8 (July 29, 2021).
21. Y. Zhang, *et al.*, Prediction of Soil Organic Carbon based on Landsat 8 Monthly NDVI Data for the Jianghan Plain in Hubei Province, China. *Remote Sens.* **11**, 1683 (2019).
22. W. Marshall, C. Boshuizen, Planet Labs' Remote Sensing Satellite System. *Small Satell. Conf.* (2013).
23. S. E. Fick, R. J. Hijmans, WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302–4315 (2017).
24. A. Cotrina Sánchez, *et al.*, Biogeographic Distribution of Cedrela spp. Genus in Peru Using

MaxEnt Modeling: A Conservation and Restoration Approach. *Diversity* **13**, 261 (2021).
25. L. Poggio, E. Simonetti, A. Gimona, Enhancing the WorldClim data set for national and regional applications. *Sci. Total Environ.* **625**, 1628–1643 (2018).
26. R. Benestad, Downscaling Climate Information. *Oxf. Res. Encycl. Clim. Sci.* (2016) https:/doi.org/10.1093/acrefore/9780190228620.013.27 (April 1, 2022).
27. M. R. Trivedi, P. M. Berry, M. D. Morecroft, T. P. Dawson, Spatial scale affects bioclimate model projections of climate change impacts on mountain plants. *Glob. Change Biol.* **14**, 1089–1103 (2008).
28. , Population Projections: Total Population for California and Counties (2021) (March 30, 2022).
29. B. J. Enquist, *et al.*, The commonness of rarity: Global and future distribution of rarity across land plants. *Sci. Adv.* **5**, 1–14 (2019).
30. N. Japkowicz, S. Stephen, The class imbalance problem: A systematic study. *Intell. Data Anal.* **6**, 429–449 (2002).
31. M. Zhu, *et al.*, Class Weights Random Forest Algorithm for Processing Class Imbalanced Medical Data. *IEEE Access* **6**, 4641–4652 (2018).
32. P. Chu, X. Bian, S. Liu, H. Ling, Feature Space Augmentation for Long-Tailed Data. *ArXiv200803673 Cs* (2020) (August 27, 2020).
33. A. T. Taylor, T. Hafen, C. T. Holley, A. González, J. M. Long, Spatial sampling bias and model complexity in stream-based species distribution models: A case study of Paddlefish (Polyodon spathula) in the Arkansas River basin, USA. *Ecol. Evol.* **10**, 705–717 (2020).
34. F. Pedregosa, *et al.*, Scikit-learn: Machine Learning in Python. *Mach. Learn. PYTHON*, 6.
35. A. Kubany, *et al.*, Comparison of state-of-the-art deep learning APIs for image multi-label classification using semantic metrics. *Expert Syst. Appl.* **161**, 113656 (2020).
36. Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
37. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2012).
38. O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation. 1–8.
39. N. Jean, *et al.*, Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790–794 (2016).
40. B. Duke, A. Ahmed, C. Wolf, P. Aarabi, G. W. Taylor, SSTVOS: Sparse Spatiotemporal Transformers for Video Object Segmentation. *ArXiv210108833 Cs* (2021) (January 27, 2021).
41. Z. Liu, *et al.*, A ConvNet for the 2020s. *ArXiv220103545 Cs* (2022) (February 4, 2022).
42. Y. S. Aurelio, G. M. de Almeida, C. L. de Castro, A. P. Braga, Learning from Imbalanced Data Sets with Weighted Cross-Entropy Function. *Neural Process. Lett.* **50**, 1937–1949 (2019).
43. B. Deneu, *et al.*, Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLoS Comput. Biol.* **17**, e1008856 (2021).
44. E. Ben-Baruch, *et al.*, Asymmetric Loss For Multi-Label Classification. *ArXiv200914119 Cs* (2020) (November 5, 2020).
45. C. Botella, A. Joly, P. Monestiez, P. Bonnet, F. Munoz, Bias in presence-only niche models related to sampling effort and species niches: Lessons for background point selection. *PLOS ONE* **15**, e0232078 (2020).
46. K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv14091556 Cs* (2015) (July 30, 2021).
47. K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition. *ArXiv151203385 Cs* (2015) (July 30, 2021).
48. , Improved Image Based Super Resolution and Concrete Crack Prediction Using Pre-Trained Deep

Learning Models (April 11, 2022).
49. T. Won, *et al.*, Particulate Matter Estimation from Public Weather Data and Closed-Circuit Television Images. *KSCE J. Civ. Eng.* **26**, 865–873 (2022).
50. , Multi-Class Weather Classification Using ResNet-18 CNN for Autonomous IoT and CPS Applications | IEEE Conference Publication | IEEE Xplore (April 11, 2022).
51. A. Demir, F. Yilmaz, O. Kose, Early detection of skin cancer using deep learning architectures: resnet-101 and inception-v3 in *2019 Medical Technologies Congress (TIPTEKNO)*, (2019), pp. 1–4.
52. T. Ridnik, *et al.*, TResNet: High Performance GPU-Dedicated Architecture. *ArXiv200313630 Cs Eess* (2020) (July 22, 2021).
53. K. He, R. Girshick, P. Dollár, Rethinking ImageNet Pre-training. *ArXiv181108883 Cs* (2018) (August 11, 2021).
54. E. Cole, X. Yang, K. Wilber, O. Mac Aodha, S. Belongie, When Does Contrastive Visual Representation Learning Work? *ArXiv210505837 Cs* (2021) (June 15, 2021).
55. S. Kornblith, J. Shlens, Q. V. Le, Do Better ImageNet Models Transfer Better? in (2019), pp. 2661–2671.
56. J. Deng, *et al.*, ImageNet: A large-scale hierarchical image database in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (2009), pp. 248–255.
57. T.-Y. Lin, *et al.*, Microsoft COCO: Common Objects in Context in *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science., D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars, Eds. (Springer International Publishing, 2014), pp. 740–755.
58. A. Norberg, *et al.*, A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecol. Monogr.* **89**, e01370 (2019).
59. J. Elith*, *et al.*, Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29**, 129–151 (2006).
60. J. Elith, *et al.*, A statistical explanation of MaxEnt for ecologists. *Divers. Distrib.* **17**, 43–57 (2011).
61. J. S. Evans, M. A. Murphy, Z. A. Holden, S. A. Cushman, "Modeling Species Distribution and Change Using Random Forest" in *Predictive Species and Habitat Modeling in Landscape Ecology: Concepts and Applications*, C. A. Drew, Y. F. Wiersma, F. Huettmann, Eds. (Springer, 2011), pp. 139–159.
62. S. J. Phillips, M. Dudík, R. E. Schapire, A maximum entropy approach to species distribution modeling in *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04., (Association for Computing Machinery, 2004), p. 83.
63. Robert J. Hijmans, Steven Phillips, John Leathwick, Jane Elith, Package 'dismo' (2021).
64. C. Merow, M. J. Smith, J. A. Silander, A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography* **36**, 1058–1069 (2013).
65. T. Kattenborn, J. Eichel, F. E. Fassnacht, Convolutional Neural Networks enable efficient, accurate and fine-grained segmentation of plant species and communities from high-resolution UAV imagery. *Sci. Rep.* **9**, 17656 (2019).
66. B. Deneu, A. Joly, P. Bonnet, M. Servajean, F. Munoz, How Do Deep Convolutional SDM Trained on Satellite Images Unravel Vegetation Ecology ? in (Springer, 2021), p. 148.
67. , Plant Range - Calflora (October 20, 2021).
68. , Calflora: Information on California plants for education, research and conservation. [web application]. Berkeley, California: The Calflora Database [a non-profit organization]. Available: https://www.calflora.org/ (Accessed: April 18, 2022).
69. Felipe Osorio, Ronny Vallejos, Francisco Cuevas, Diego Mancilla, SpatialPack (2020).
70. A. Siefert, C. Ravenscroft, M. D. Weiser, N. G. Swenson, Functional beta-diversity patterns reveal

deterministic community assembly processes in eastern North American trees. *Glob. Ecol. Biogeogr.* **22**, 682–691 (2013).
71. M. Večeřa, *et al.*, Alpha diversity of vascular plants in European forests. *J. Biogeogr.* **46**, 1919–1935 (2019).
72. Y. Yang, D. Tilman, G. Furey, C. Lehman, Soil carbon sequestration accelerated by restoration of grassland biodiversity. *Nat. Commun.* **10**, 718 (2019).
73. T. H. Oliver, *et al.*, Biodiversity and Resilience of Ecosystem Functions. *Trends Ecol. Evol.* **30**, 673–684 (2015).
74. B. Sakschewski, *et al.*, Resilience of Amazon forests emerges from plant trait diversity. *Nat. Clim. Change* **6**, 1032–1036 (2016).
75. C. R. Lawson, J. A. Hodgson, R. J. Wilson, S. A. Richards, Prevalence, thresholds and the performance of presence–absence models. *Methods Ecol. Evol.* **5**, 54–64 (2014).
76. E. A. Freeman, G. G. Moisen, A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecol. Model.* **217**, 48–58 (2008).
77. Marin GIS Admin, Marin Parks (2017) (February 12, 2022).
78. Marin GIS Admin, Ridge and Upland Greenbelt (2017) (February 12, 2022).
79. QGIS Association, QGIS Geographic Information System (2022) (April 19, 2022).
80. California State Parks, State Park Boundaries of California (2022) (February 12, 2022).
81. San Francisco City Planning, San Francisco Land Use (February 12, 2022).
82. OpenStreetMap Foundation, OpenStreetMap BaseMap (2022) (February 12, 2022).
83. GADM, Global Administrative Areas (GADM) (2022) (February 12, 2022).
84. Kelsey Jordahl, *et al.*, GeoPandas (2020) (April 25, 2022).
85. A. Marathe, D. R. Priyadarsanan, J. Krishnaswamy, K. Shanker, Gamma diversity and under-sampling together generate patterns in beta-diversity. *Sci. Rep.* **11**, 21420 (2021).
86. Golden Gate National Parks Conservancy, Tamalpais Lands Collaborative (One Tam), Aerial Information Systems, Tukman Geospatial LLC, Marin County Fine Scale Vegetation Map (2021) (February 14, 2022).
87. Sean Gillies, *et al.*, Shapely (2021) (April 25, 2022).
88. H. C. Chandler, J. C. Colón-Gaud, T. A. Gorman, K. Carson, C. A. Haas, Does long-term fire suppression impact leaf litter breakdown and aquatic invertebrate colonization in pine flatwoods wetlands? *PeerJ* **9**, e12534 (2021).
89. Z. L. Steel, H. D. Safford, J. H. Viers, The fire frequency-severity relationship and the legacy of fire suppression in California forests. *Ecosphere* **6**, art8 (2015).
90. E. N. Stavros, *et al.*, Remote Sensing Data Before and After California Rim and King Forest Fires, 2010-2015. *ORNL DAAC* (2016) https:/doi.org/10.3334/ORNLDAAC/1288 (January 28, 2022).
91. J. E. Keeley, Fire intensity, fire severity and burn severity: A brief review and suggested usage. *Int. J. Wildland Fire* **18**, 11 (2009).
92. P. Dutilleul, P. Clifford, S. Richardson, D. Hemon, Modifying the t Test for Assessing the Correlation Between Two Spatial Processes. *Biometrics* **49**, 305 (1993).
93. Laurel Anders, Pond It and Plug It: Restoring Wet Meadows in Northern California. *Intermt. West Jt. Venture* (2013) (April 25, 2022).
94. K. Morefield, Ash Creek Wildlife Refuge boundary and ponds, CDFW (2012). (April 25, 2022).
95. Lisa Ohara, Owned and Operated Lands and Conservation Easements (2021) (April 25, 2022).
96. C. R. Keyes, E. K. Teraoka, Structure and Composition of Old-Growth and Unmanaged Second-Growth Riparian Forests at Redwood National Park, USA. *Forests* **5**, 256–268 (2014).
97. GBIF.org (12 May 2022) GBIF Occurrence Download https://doi.org/10.15468/dl.gt624q
98. B. Baldwin, A. Thornhill, W. A. Freyman, D. D. Ackerly, M. M. Kling, N. Morueta-Holme, and B. D. Mishler. Species richness and endemism in the native flora of California. *American Journal*

of Botany **104**, 487–501 (2017).

99. Nelson, Brewer, and Solem. Existing Vegetation Classification, Mapping, and Inventory Technical Guide EVeg Mid Region 5 South. U.S. Forest Service, January 18, 2018.https://data.fs.usda.gov/geodata/edw/edw_resources/meta/S_USA.EVMid_R05_SouCoast.xml
100. South Coast and Montane Ecological Province Vegetation Description. U.S. Forest Service.https://www.fs.usda.gov/Internet/FSE_DOCUMENTS/fsbdev3_045405.pdf
101. R. Valavi, G. Guillera-Arroita, J. J. Lahoz-Monfort, J. Elith, Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecological Monographs* **92** (2022).

SUPPLEMENTAL FIGURES

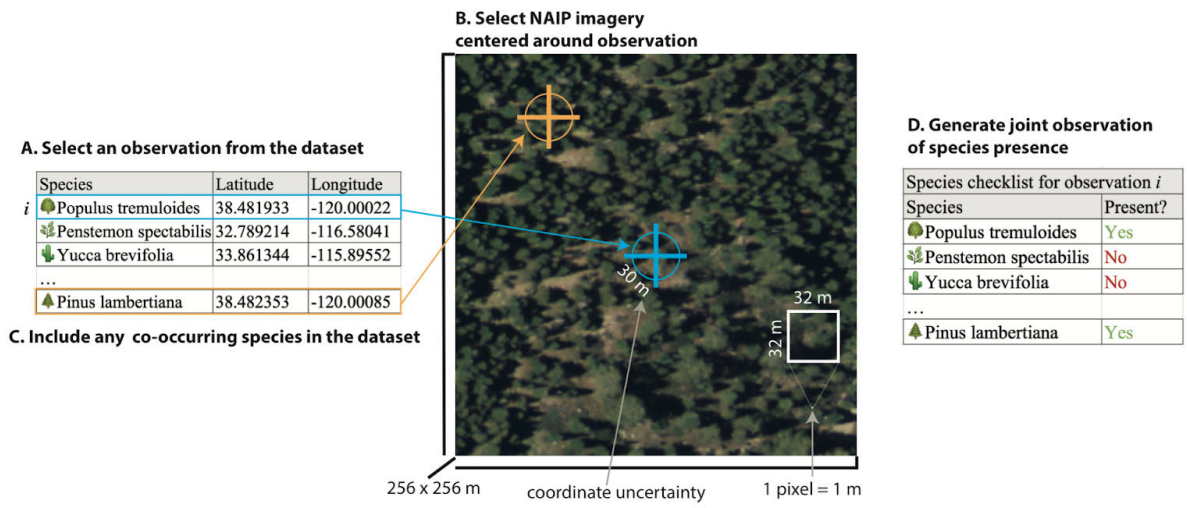

Fig. S1 | Illustration of joint dataset construction.
For every unique observation in the dataset, a 256 × 256 pixel image (pixel resolution is 1m) is generated from the NAIP imagery (B). Next, a list of overlapping species is generated by selecting other observations from the dataset whose coordinates fall within the 256m radius of the original observation (C). This gives us the final input and output pair that will be used to train the neural network. Specifically, the image (B) is the input and the species checklist (D) is the output.

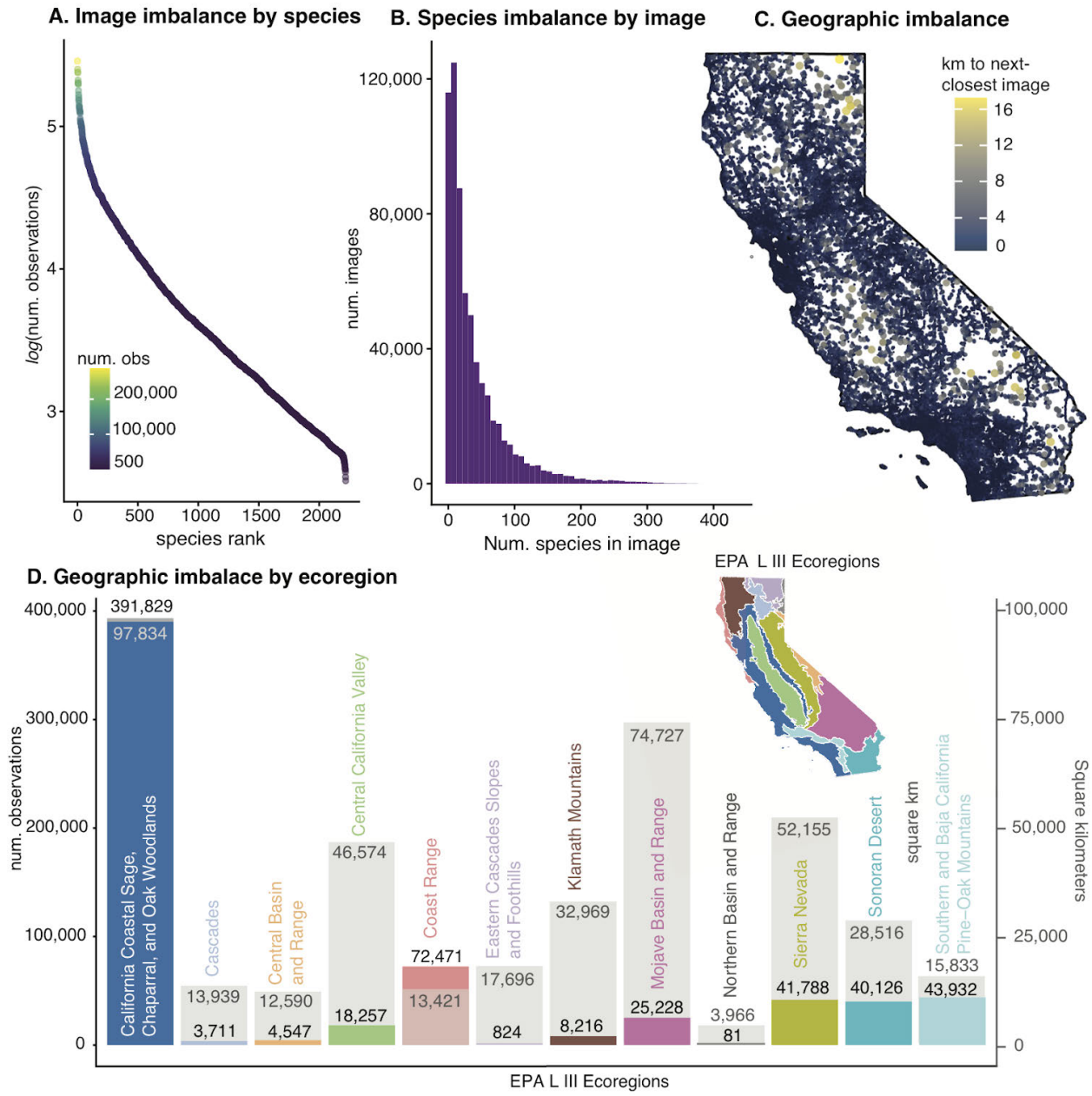

Fig. S2 | Biases present in the dataset.

(A) Many species have few images in the dataset while a few species have a disproportionate number of observations. This extreme imbalance between species class frequencies can make it difficult for classic machine learning methods – that rely on an assumption of an equal number of examples per-class – to learn a good representation. (B) This histogram highlights the significant imbalance of number of species labeled per-image in the dataset. Most observations have few overlapping species in the observation – likely not fully describing all species actually present in a given site – while only a few observations contain many species and present a more accurate checklist of species presence in a given area. (C) Plot of distance to nearest non-overlapping observation across California. Color and size represent the distance in kilometers to the next-closest non-overlapping image in the dataset. As can clearly be seen, observations tend to cluster and distances are not distributed evenly. (D) Comparison of number of species occurrences in a given Level III EPA ecoregion versus the area of

that ecoregion. Citizen science observations like those used in the dataset tend to cluster around population areas and natural regions where observers can reach. This leads to oversampling in some ecoregions, especially the California coastal sage, chaparral and oak woodlands and coast range, and significant undersampling in other regions, especially the cascades and basin and range habitats. Colored bars represent the number of observations in the dataset from that region and the grey bars represent the area in square kilometers of the specific ecoregion. Inset is a map of the level III ecoregions of California.

Fig. S3 | Comparison of different possible remote sensing products.
(A). NAIP imagery from dataset. (B). Image at same location and extent from Landsat 7 satellite imagery acquired August 27th, 2012 (Landsat-7 image courtesy of the U.S. Geological Survey). Although the pixel color is distinguishable between the forested and scrubby areas, the resolution is too coarse to distinguish local landscape patterns and individual tree crowns are indistinguishable. (C). The annual mean temperature in 2000 from WorldClim (23). WorldClim at 1 km resolution is the base product used to generate the BioClim variables often used to generate SDMs. This product also lacks the resolution to distinguish both topographic and land use factors.

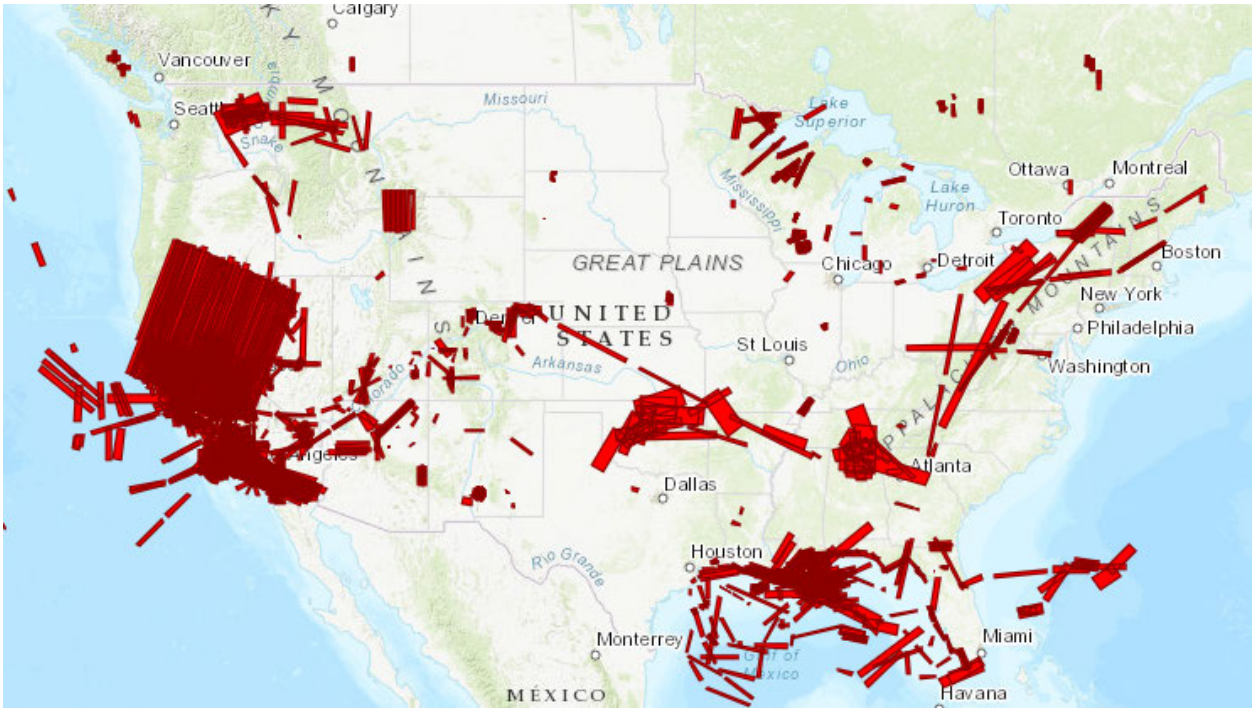

Fig. S4 | Extent of AVIRIS hyperspectral data in continental US.
Extent of AVIRIS 224 contiguous band hyperspectral collection across the continental US. Image collected from https://aviris.jpl.nasa.gov/dataportal/ on March 29th, 2022. (Image courtesy of NASA JPL data portal)

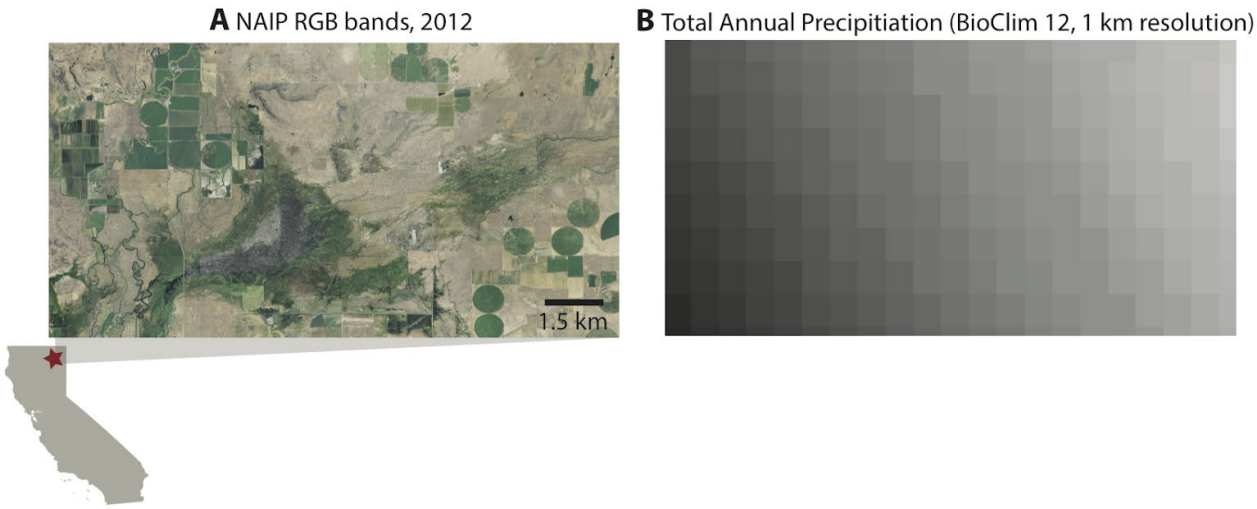

Fig. S5 | Comparison of BioClim and NAIP resolution.
Here we can compare the resolution and detectable data from remotely sensed data and BioClim variables at Ash Creek State Wildlife Area. (A) using the NIAP imagery, different ecological features of this example wetland area are clearly visible, including the marshy, darker regions at the center of the image along with grassland incursion in the center-right. (B), the 12th bioclim variable, total annual precipitation, clearly does not capture these more local differences that are clearly visible from remote sensing imagery. It should be noted that the projection for the BioClim variable is WSG84, a geographic coordinate system which does not preserve distance, thus each represented pixel is not a fully 1x1 km.

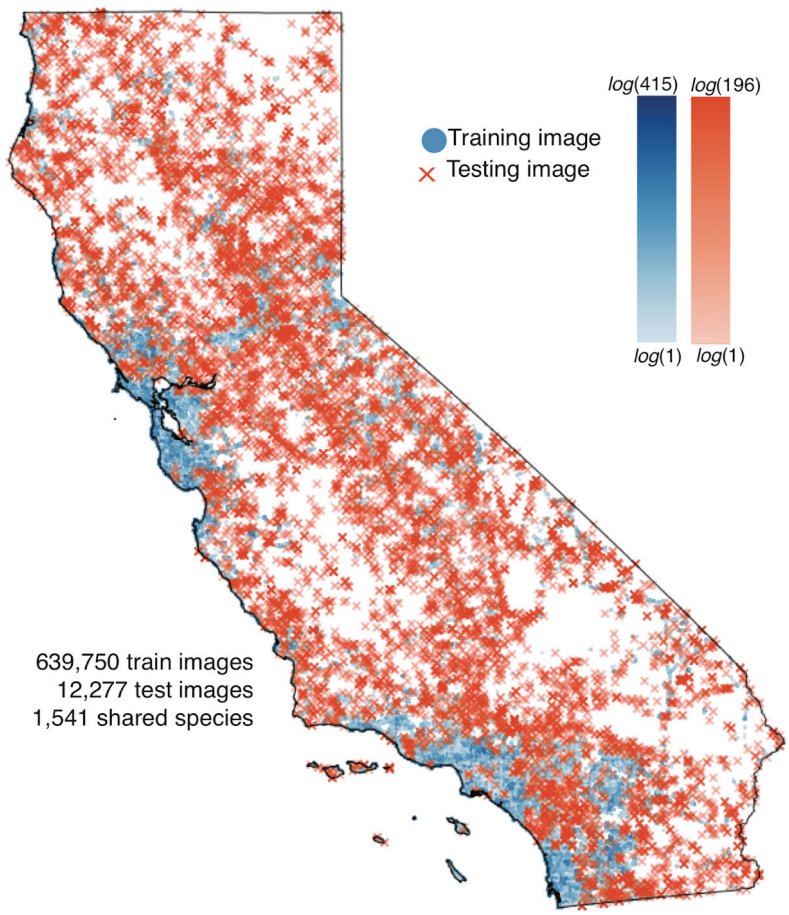

Fig. S6 | Location of train and test images used for interpolation validation
For determining the best loss function and learning rate, we generated a train/test split meant to maximize the number of ecosystems and shared species. Clusters of observations that were at least 1.3km away from any other observation were chosen as test images, with all other observations used as images for training. A cluster is defined as any group of observations where each observation is within at least 256m of another observation in the cluster.

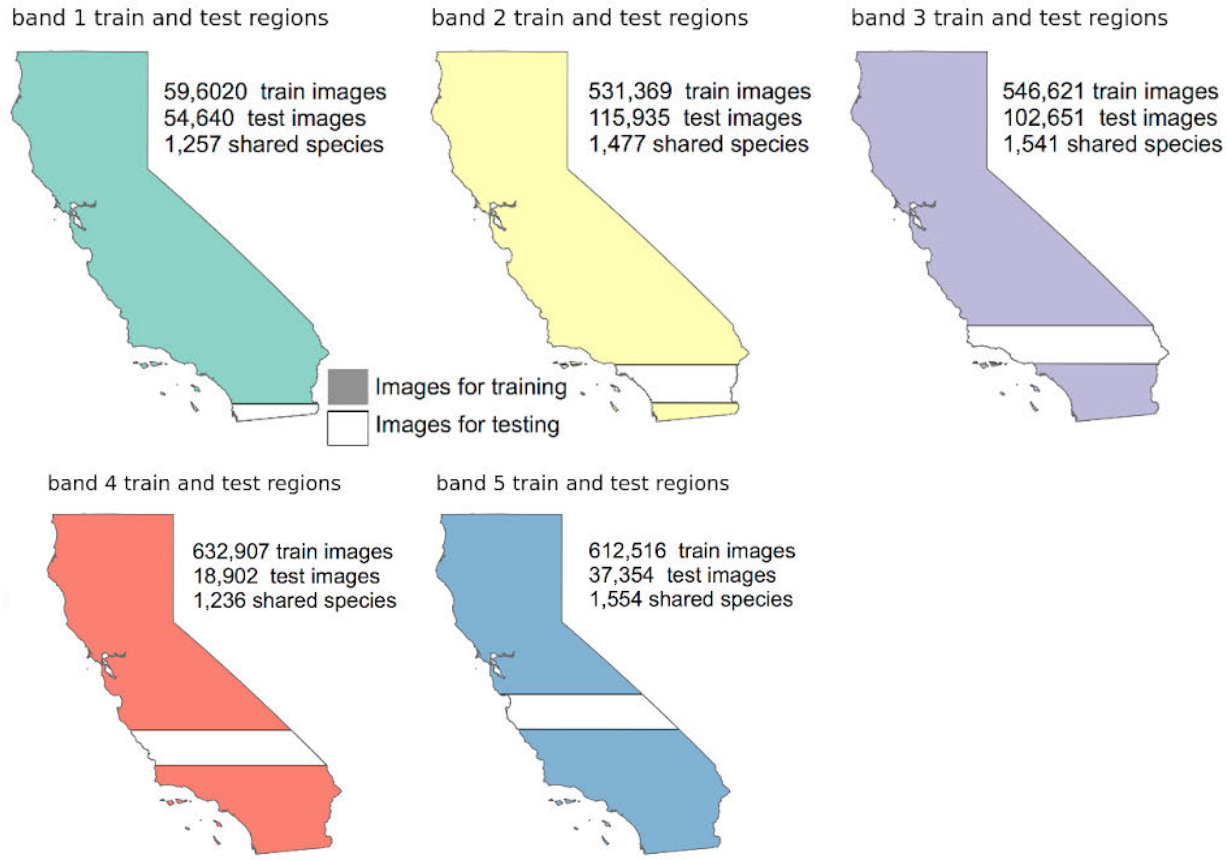

Fig. S7 | Location of splits used for spatial cross-validation

We employed a 5-block spatial holdout procedure to test the SDM's extrapolation ability. Areas that were included in the training set are in color and areas used for the test set are in white. Any images within 1.3km of the start of the test region were removed from the training set in order to prevent potential overlap between bioclimatic variables in the train and test splits. The number of images within each split is also annotated, along with the number of unique species present in both splits.

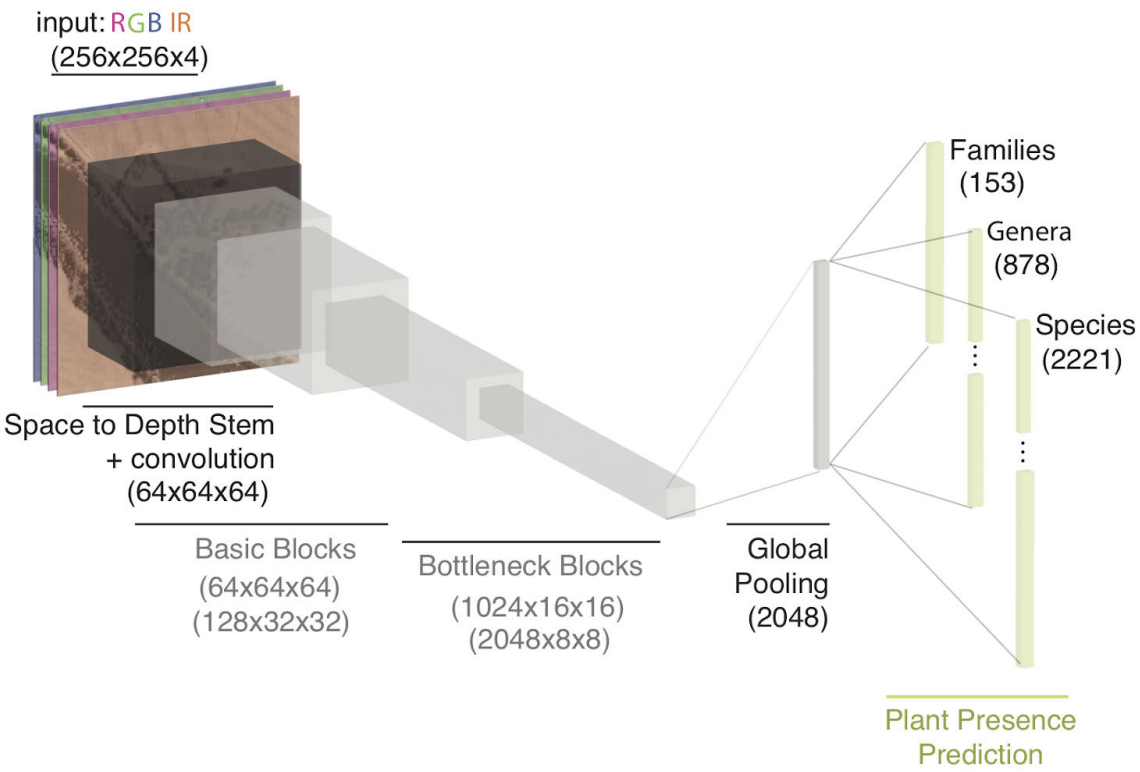

Fig. S8 | Graphical description of image-only TResNet convolutional neural network (CNN) architecture

This CNN uses only the ultra-high-resolution NAIP aerial imagery to make predictions about what plant species are present in a given image. The preliminary model *deepbiosphere*$_{alpha}$ uses this architecture with a slightly modified output layer, predicting simultaneously 2,242 species, 886 genera and 181 families. The difference in output layer is because the preliminary dataset and main dataset used different methods for determining the threshold for number of occurrences a species should have to be included in the dataset.

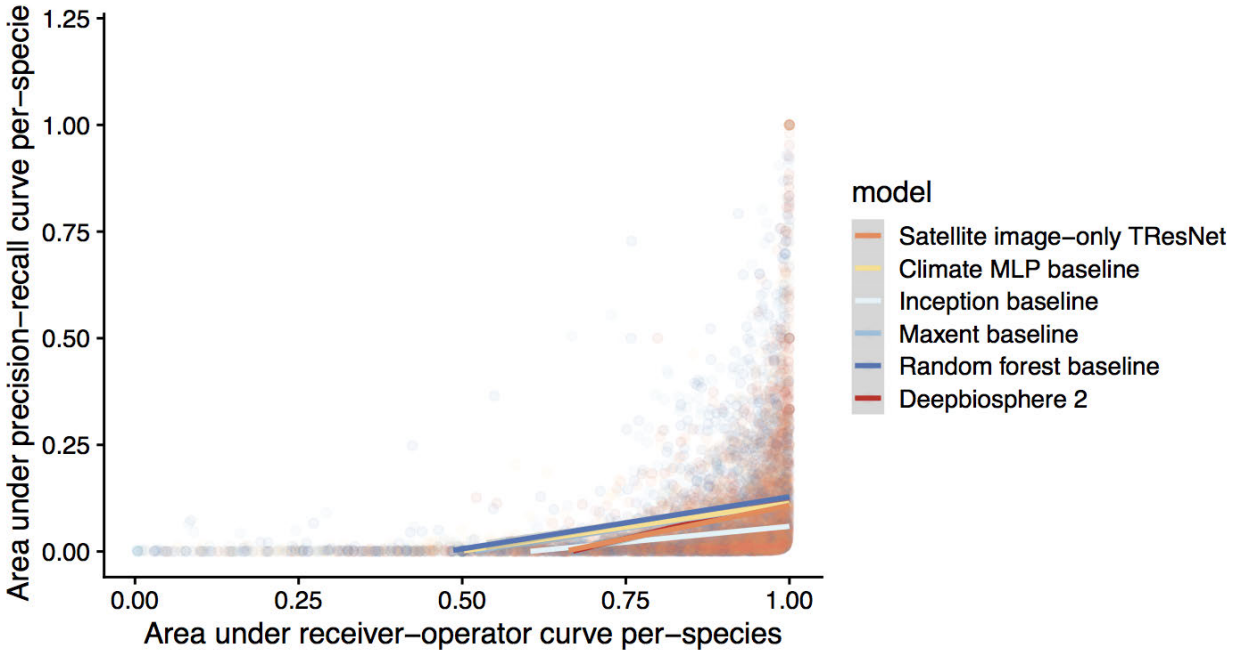

Fig. S9 | Relationship of area under the precision-recall curve with area under the receiver-operator curve per-species

Each dot represents one of the 1,542 species shared between the uniform train and test split of the dataset. Lines of best fit are linear models fitted using base R's linear model function. A larger x intercept is better, indicating said model tends to have a higher AUC_{ROC} per-species, and a steeper slope is also better, as it indicates that AUC_{PRC} tends to be higher when the AUC_{ROC} is also high.

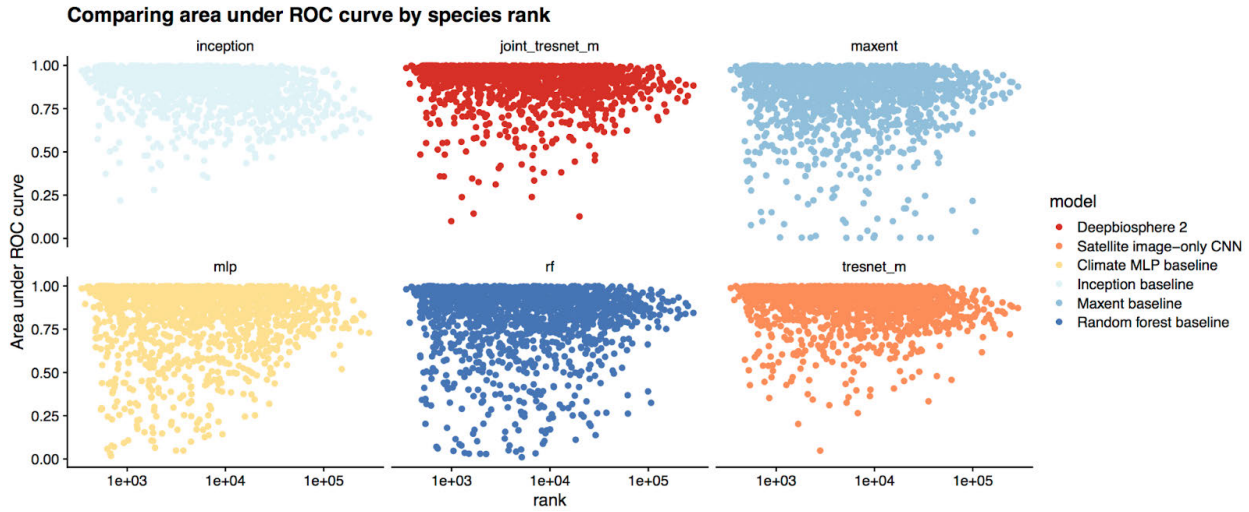

Fig. S10 | Comparing area under the receiver-operator characteristic curve by species rank
Each dot represents one of the 1,542 species shared between the uniform train and test split of the dataset. The x axis is $\log_{10}$ scaled. Higher areas as rank increases is better, as it indicates that rarer species still tend to have a high AUC_{ROC}.

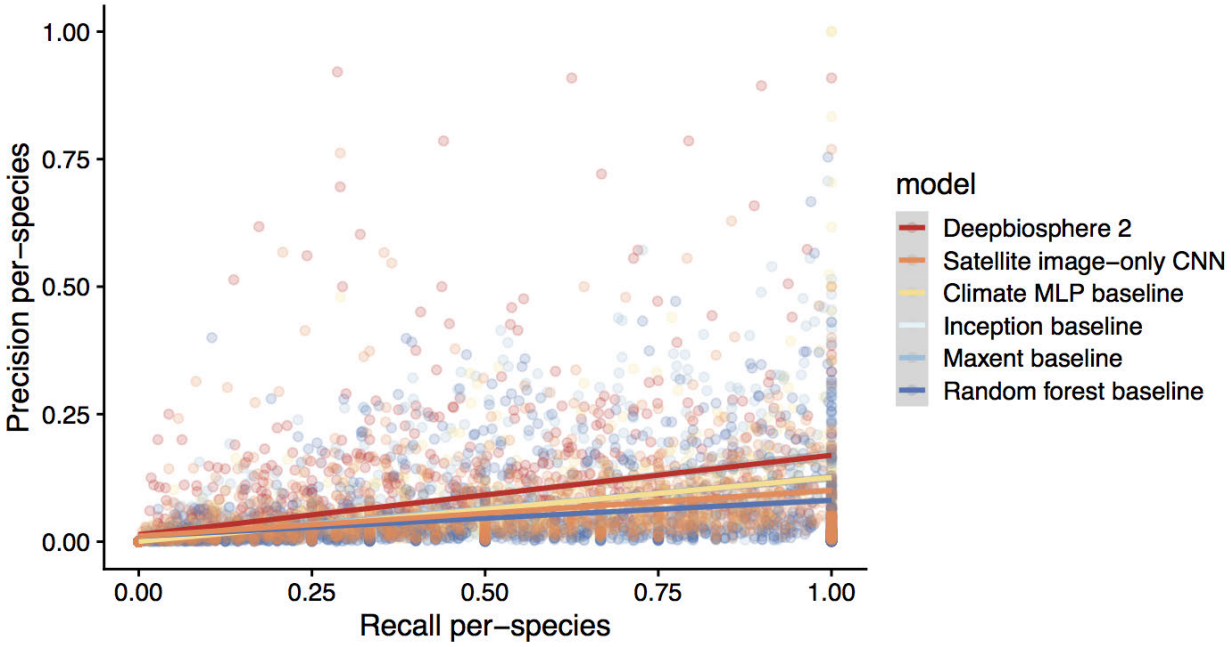

Fig. S11 | Tradeoff between recall and precision per-species across models
Each dot represents one of the 1,542 species shared between the uniform train and test split of the dataset. Lines of best fit are linear models fitted using base R's linear model function. Larger area under the curve indicates that the model tends to have a higher precision when the recall is higher.

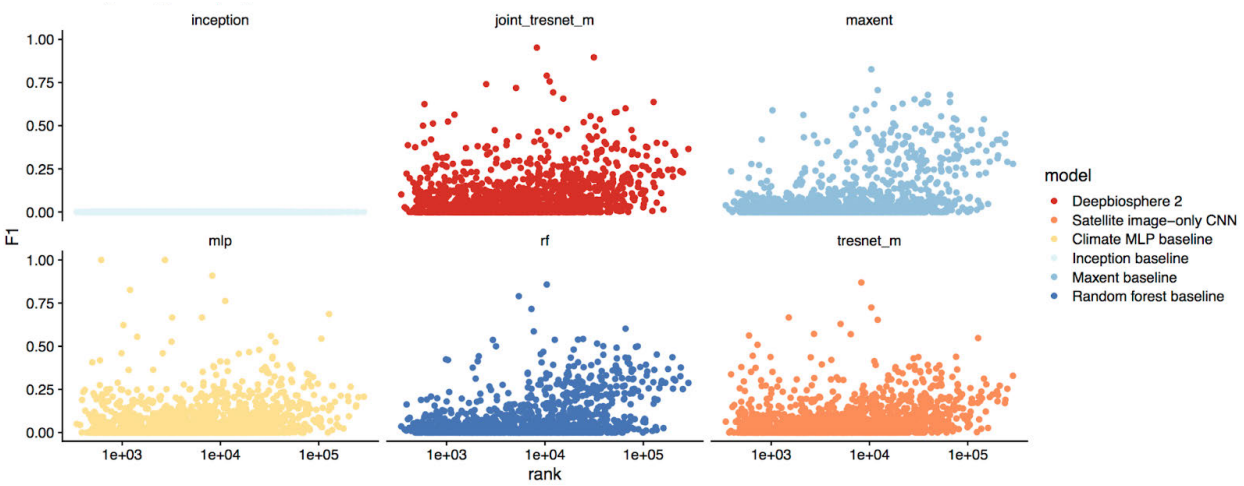

Fig. S12 | Comparing F1 by species rank

Each dot represents one of the 1,542 species shared between the uniform train and test split of the dataset. The x axis is $\log_{10}$ scaled. Higher F1 as rank increases is better, as it indicates that rarer species still tend to have accuracy when using the binary classification metrics of recall and precision.

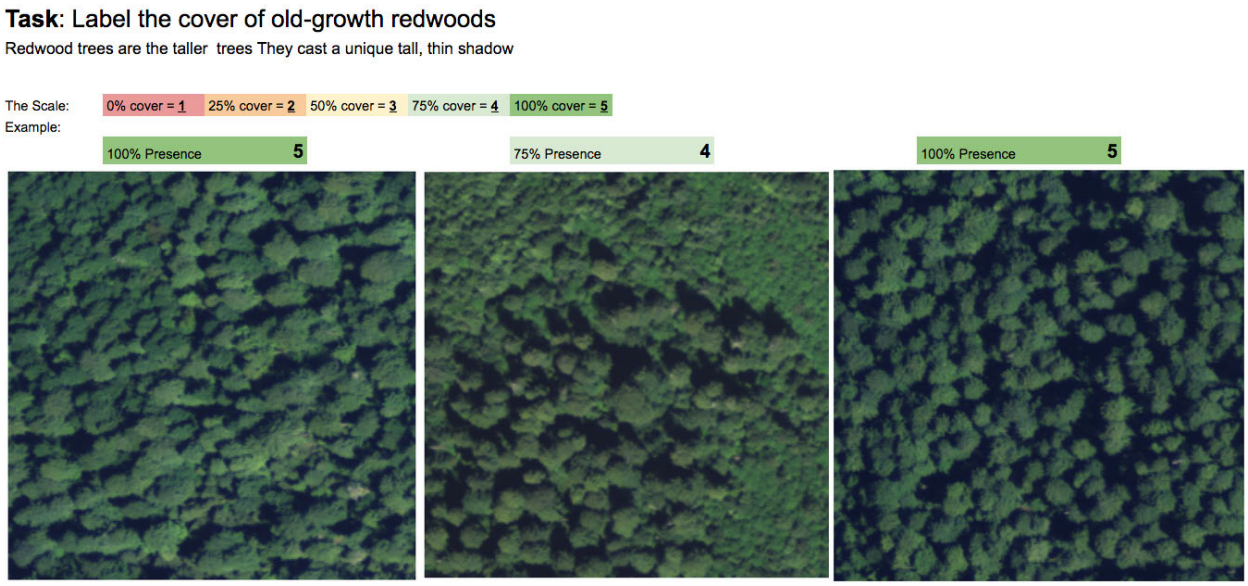

Fig. S13 | Example images from human redwood labeling task.
Labelers were given these three examples of old-growth redwoods as positive examples of redwood forest. Details of the three locations can be found in **Table S5**.

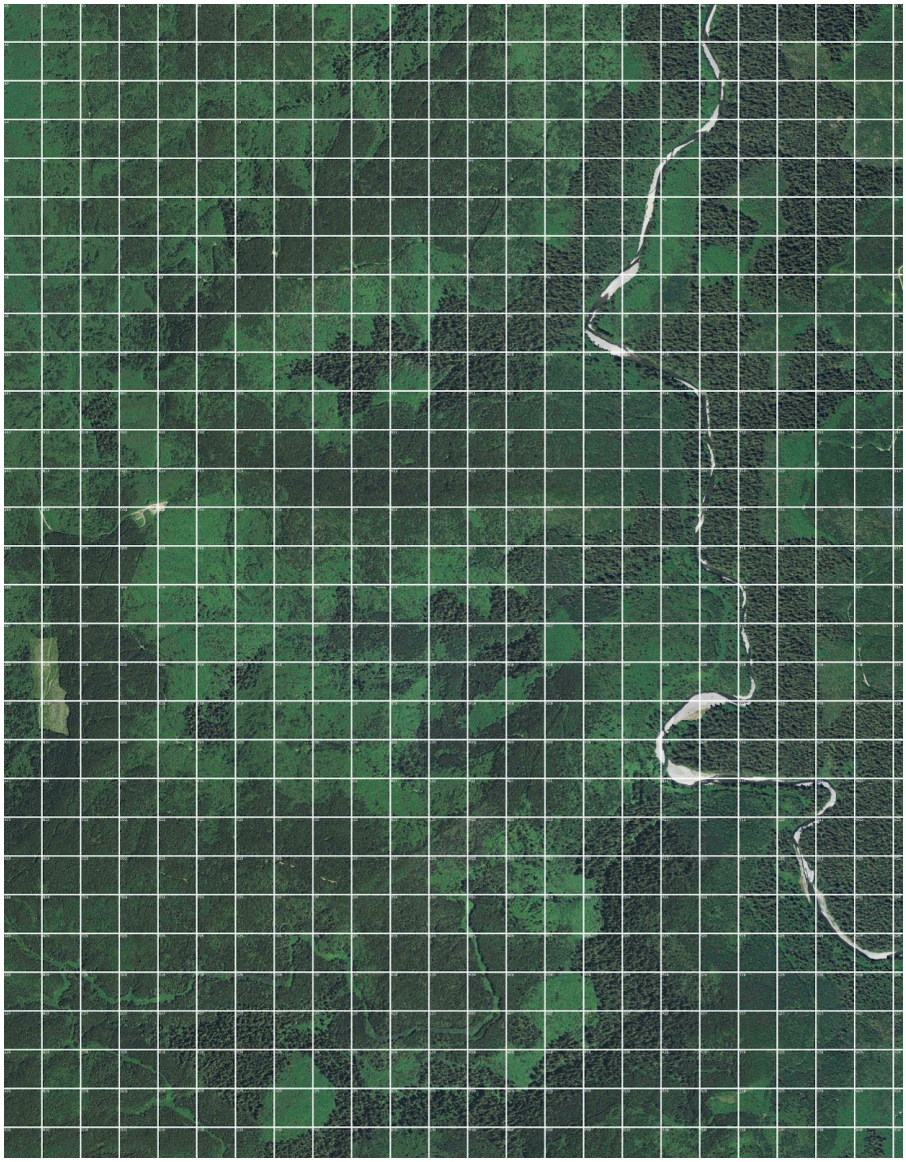

Fig. S14 | Aerial imagery used for redwoods human annotation task.
Cells are 256 × 256m in size, to map to the same resolution as *deepbiosphere's* predictions. Each cell was annotated [A-Z][1-30] to correspond with cells in Google Sheets. For each cell, labelers annotated their estimation of the percent old-growth redwood cover using the examples from **Fig. S13** as guide.

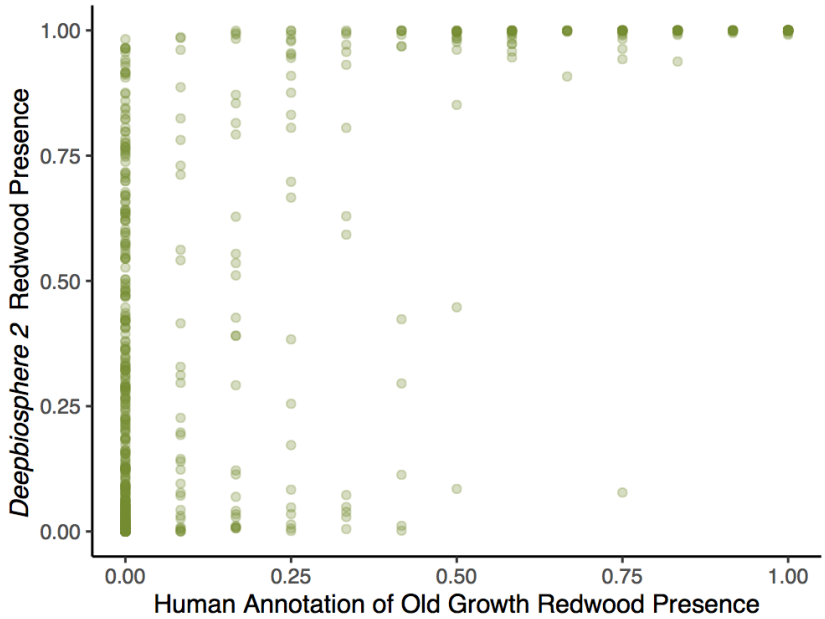

Fig. S15 | Correlation of human annotations and *deepbiosphere's* predictions
Comparison of human annotations of redwood presence compared to *deepbiosphere*'s predicted redwood presence per-cell. Pearson's R: 0.728, P-value < 0.0005; Spearman's correlation: 0.689, P-value < 0.0005; Cohen's Kappa: 0.251. Human annotation scores are averaged across three independent annotators. Cohen's Kappa calculated by binning probabilities and annotations into five bins, to match the original annotation scores. While generally annotations between the model and the humans agree, there are some cells which *deepbiosphere* is quite confident are redwoods while human annotators disagree. It's plausible that these cells represent areas of secondary re-growth of the previously clear-cut forest as secondary growth redwood forests that grow back without thinning from clear-cut stands tend to be more dense with narrower, shorter trees with more canopy crowding than typically present in old-growth stands (96).

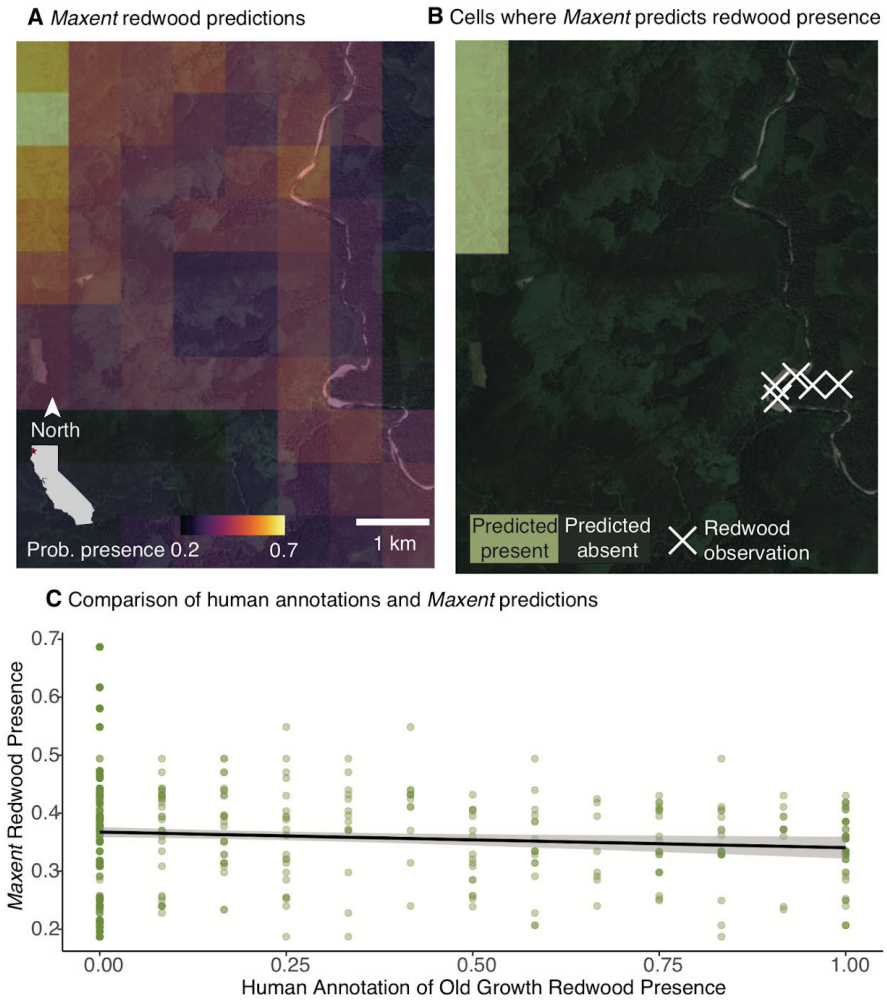

Fig. S16 | *MaxEnt* redwoods presence prediction

(A) *MaxEnt* prediction of *Sequoia sempervirens* species distribution from coarse bioclimatic correlates (Recall =.384). Color intensity maps to predicted probability at each kilometer pixel. (B) Localisation of which cells are predicted as containing and not containing redwoods according to *MaxEnt* compared to known locations of redwoods, using a standard threshold of 0.5 for presence. As can be seen, *MaxEnt* does not predict the cells containing known redwood images from the training dataset. (C) Correlation of MaxEnt prediction of *Sequoia sempervirens* species presence and human annotation of old-growth Redwood presence per-cell. There is not a significant correlation between *MaxEnt* predicted presence and human labels using Pearson's r correcting for spatial autocorrelation using the Dutilleul test, (r: -.092, P value = 0.636) and Spearman's rho (ϱ:-.069, P value =.072). Annotator's agreement is also low with Cohen's kappa (kappa = .030).

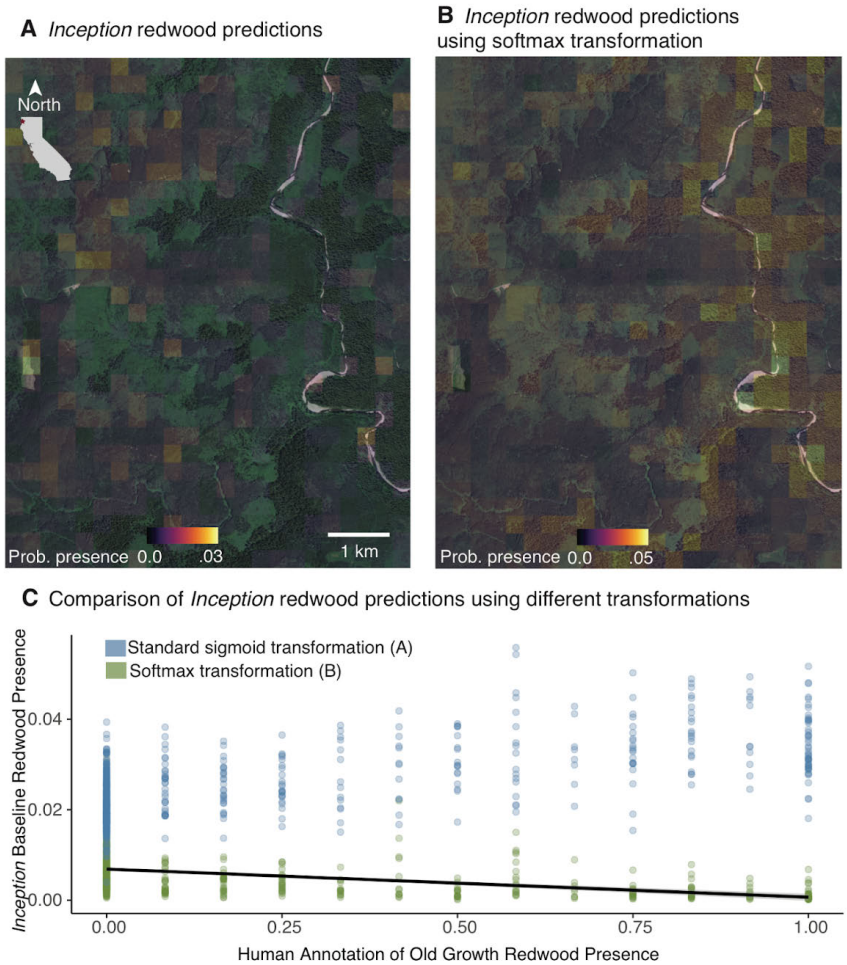

Fig. S17 | Inception V3 baseline redwoods presence prediction

(A) *Inception* baseline CNN SDM's prediction of *Sequoia sempervirens* species distribution from remote sensing imagery using standard sigmoid probability transformation (Recall = 0.0). While *Inception* has the same resolution as *deepbiosphere*, its single-label nature which ignores overlapping species information and its cross-entropy based loss function that is meant to match a probability distribution across classes makes it a poor choice for use as a joint species distribution model. When using the standard sigmoid transformation for interpreting the CNN's outputs as probabilities –which allows probabilities to be compared between species– nowhere in the image does *Inception* predict above the standard 0.5 presence threshold, and nor does it seem to be detecting the old-growth groves. (B) If we instead use the softmax transformation which cross-entropy loss trains to fit, we do see that the CNN's predictions become more reasonable, but still at too low of a value to be interpreted as a probability of presence index, as all predicted values are still far below the 0.5 presence threshold. Furthermore, using the softmax transformation means that the predicted probabilities for other species are no longer independent, and presence maps cannot be reliably built for multiple species at a time, a key prerequisite for a joint SDM (**SM 3.2.5**). (C) As evidenced in (A), when using the correct sigmoid transformation, *Inception*'s predicted probabilities are inversely correlated with human annotations of redwood cover (Pearson's r = -0.46, P-value < 10^{-4} with Dutilleul test), not a good behavior for an SDM that's meant to predict redwood presence. When using the less correct softmax transformation, the probabilities are positively correlated (Pearson's r = 0.66, P-value < 10^{-5} with Dutilleul test) but the inter-annotator agreement is zero (Cohen's kappa = 0).

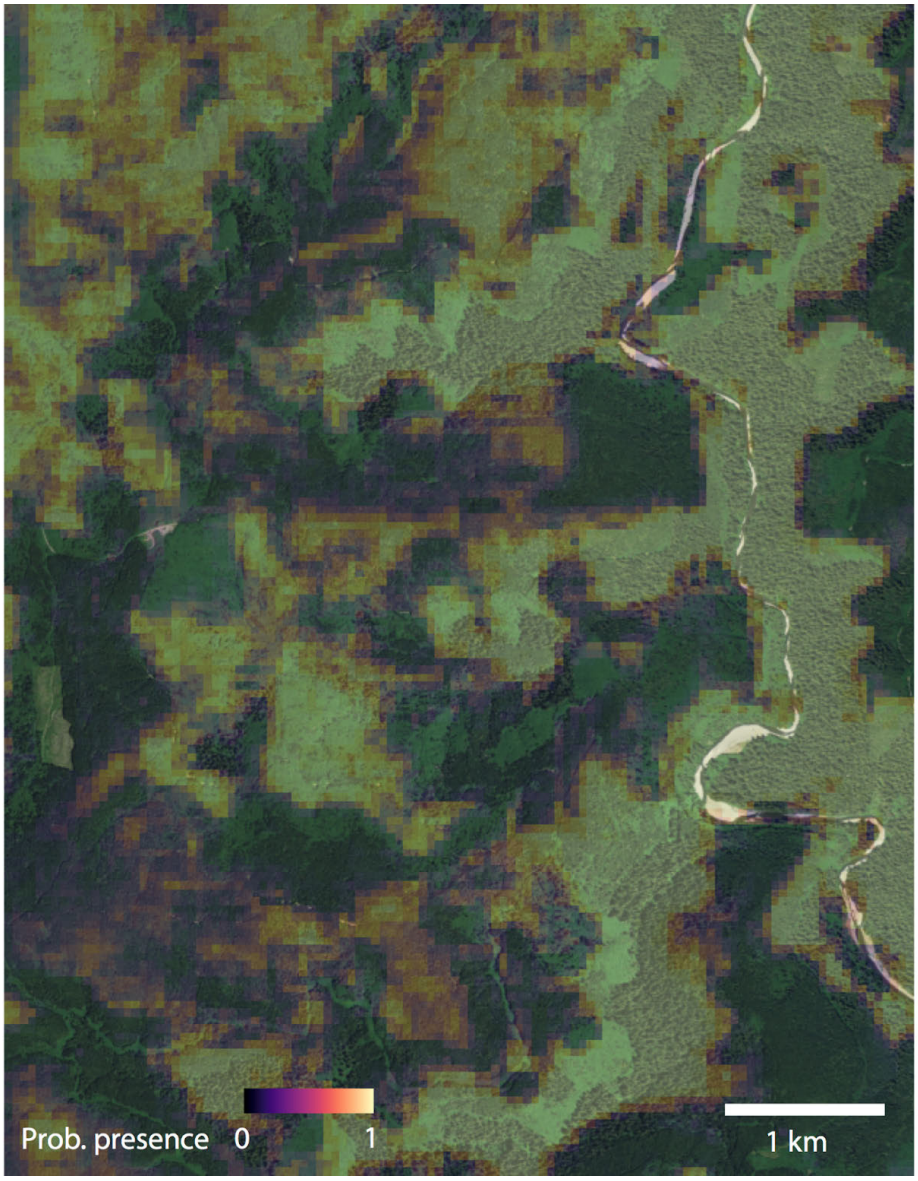

Fig. S18 | *Deepbiosphere* redwood predicted presence at 50m resolution
A benefit of our approach is that thanks to the convolutional nature of CNNs, predictions can be made up to the resolution of the underlying remote sensing imagery used to train the networks –in this case 1m – by convolving the model pixel by pixel. Each pixel maps to the predictions of a 256 × 256m image starting with said pixel as the top left pixel. This simplistic approach means predictions are slightly shifted to the top left, but more sophisticated methods exist for centering predictions to their underlying pixel to counteract this effect.

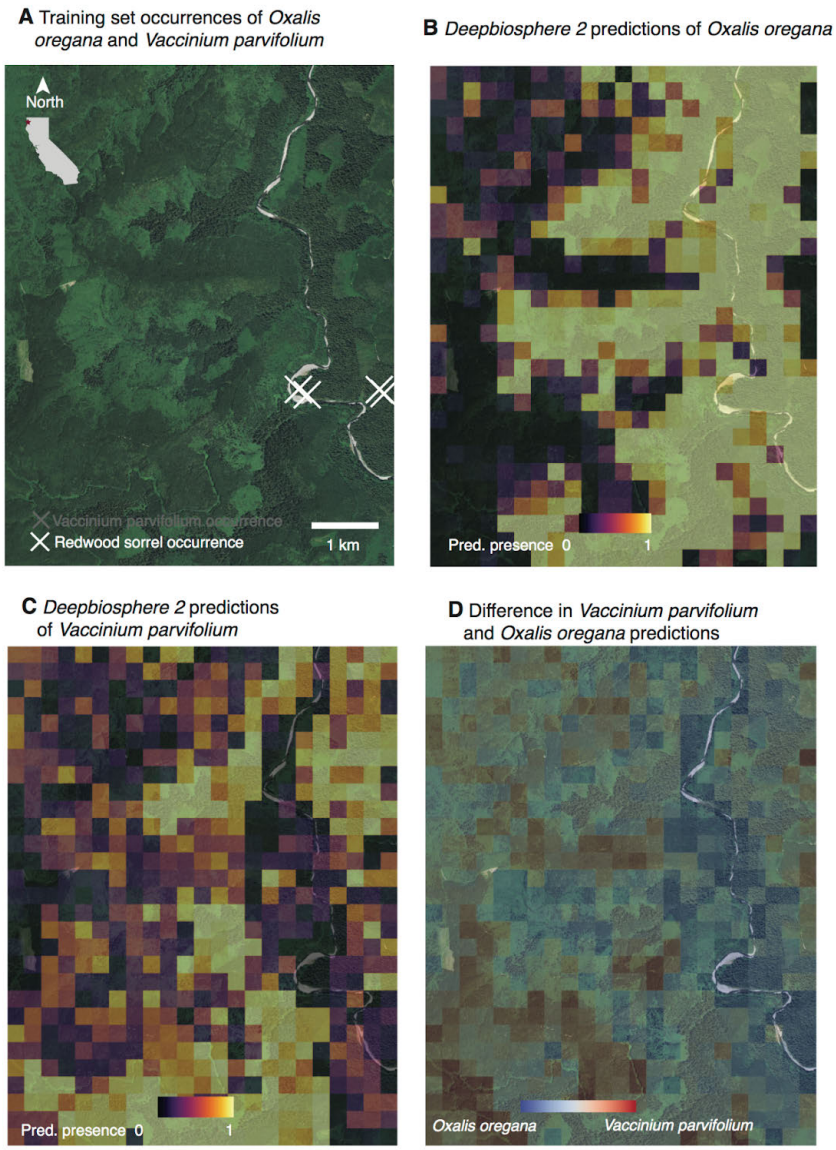

Fig. S19 | *deepbiosphere* captures interesting co-occurrence patterns in small herbaceous and shrubby plants.
(A) Location of known occurrences of *Oxalis oregana* in the training set. No examples of *Vaccinium parvifolium* are present from the training set in the image. (B) Predicted presence of *Oxalis oregana* across the case study area. Predictions largely follow the same river drainage that human annotators indicated as old-growth redwood forests, with a slightly extended footprint. (C) Predicted presence of *Vaccinium parvifolium* across the case study area. Predictions are roughly inverted to that of *Oxalis oregana*, seeming to pick up more clear-cut areas than intact stands. (D) Comparing the difference in *Oxalis oregana* versus *Vaccinium parvifolium* predictions. Taking the difference of maps (C) and (B), we can contrast areas where *Vaccinium parvifolium* is more highly predicted –and thus potentially detecting newer, secondary growth stands returning after mid-century clearcutting– to those where *Oxalis oregana* is more highly predicted, and likely older growth.

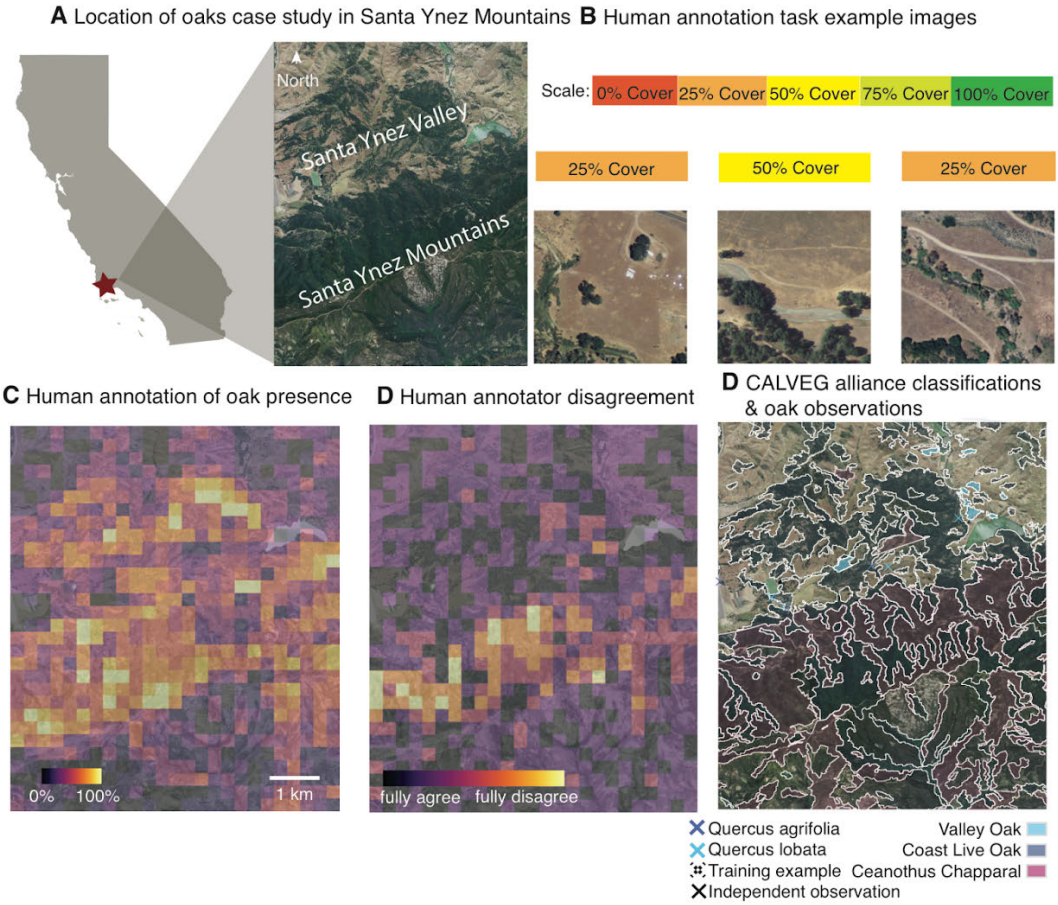

Fig. S20 | Predictions of dominant species in the Santa Ynez mountains of Southwest California. (A) The Santa Ynez mountains of Southern California have a rich and varied ecological landscape, with many rapid ecosystem transitions from sheltered but dry interior valley savanna to scrubby mountain peaks bombarded by the harsh sea breeze. This area makes for a good case study to explore how well our approach can capture subtle but important variations between adjacent ecosystems, specifically by comparing our approach's ability to distinguish Valley Oak habitat from Coast Live Oak and Buckbrush. (B). As a proxy for ground-truth annotations, we again had two human annotators label the oak cover across the region. We gave annotators these three images of known *Q. lobata* presence and the expected cover before they performed the labeling task. (C) While humans were able to successfully annotate oak presence (average human presence = 0.5), these annotations do not capture the subtle variation between the two oaks and other woody, scrubby species like *Ceanothus*. (D) The difficulty of this task is underscored by the amount of disagreement between the two annotator's annotations, with 100% disagreement in some cells. (E) Map of known presence locations of *Q. lobata* and *Q. agrifolia* from both the training set and independent observations. Independent observations were acquired from Calflora.org. The associated CALVEG vegetation mapping where our three target species are considered dominant species is also overlaid. The transition between *C. cuneatus* and *Q. agrifolia* along the spine of the Santa Ynez is jagged and by appearance not well captured by human annotations of oak cover.

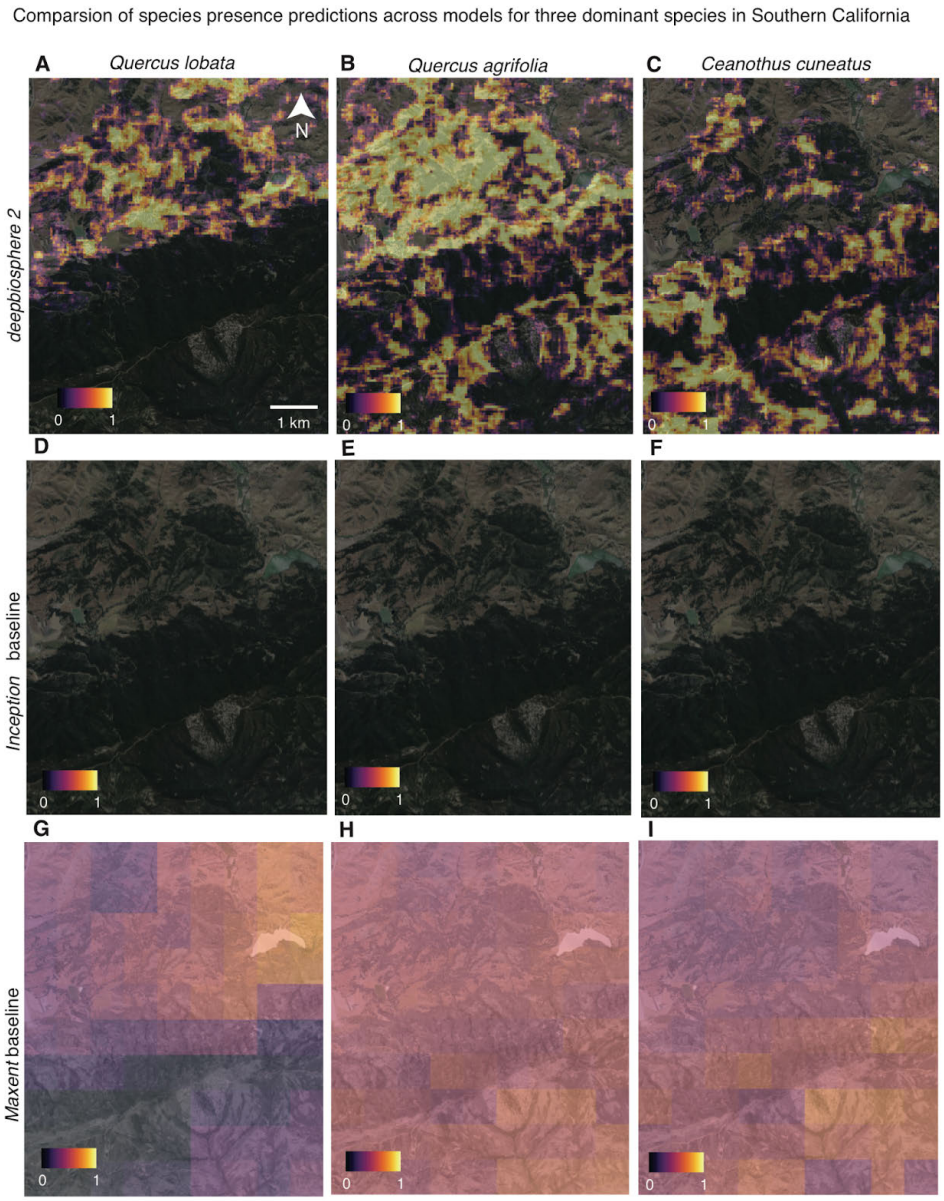

Fig. S21 | Comparison of species presence predictions across three species distribution models in Southern California

(A-C) Here we can see from *deepbiosphere* the predicted presence for the three case study species, *Q. lobata*, *Q. agrifolia*, and *C. cuneatus*. The model does predict all species as present in the study area, and in regions that roughly align with the vegetation map of Fig. S26E. Model predictions were made at 50m resolution to better capture the rapid transitions evident on the north side of the Santa Ynez mountains. (D-F) Conversely, the *Inception* baseline model transformed with the standard sigmoid transformation predicts neither species as present across the map. Predictions are shown at 50m resolution for comparison. (G-I) The MaxEnt baseline model does predict all species as present within the case study area, but clearly lacks the resolution necessary to capture the rapid variation between the three species' dominant habitats, according to the CALVEG mapping in Fig. S25E.

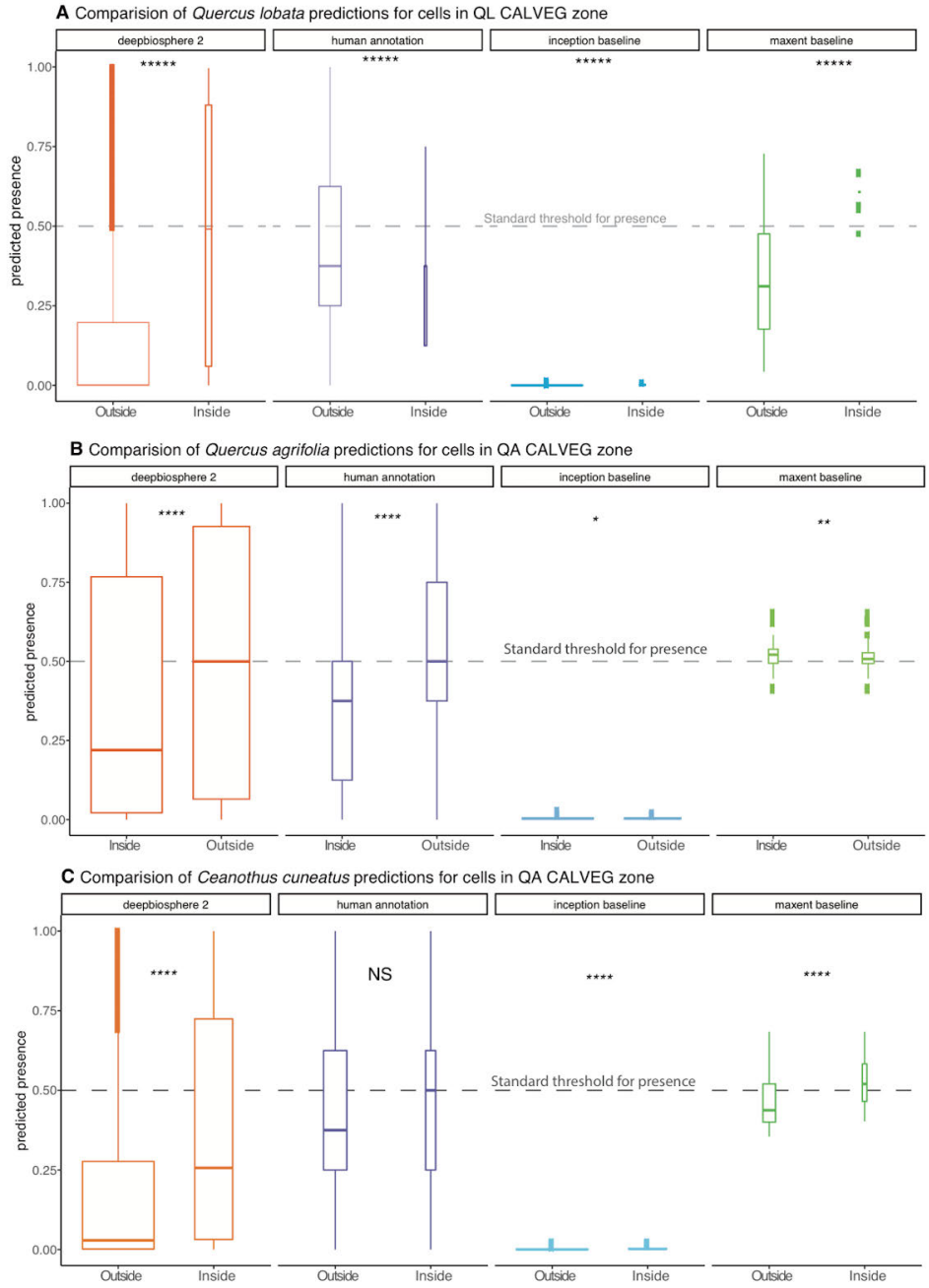

Fig. S22 | Comparison of predicted presence of case study species to CALVEG vegetation mapping.

(A) For case study species *Q. lobata*, we classified each pixel in the predictions from S25C, S26A,D,G based on whether they were inside or outside of the regions annotated as Valley Oak alliance (QL) from the CALVEG vegetation mapping in S25D. Since the area annotated as QL was quite small, we also annotated every overlapping cell with a known presence of *Q. lobata* as within the QL region. We then compared the predicted probabilities of *Q. lobata* presence from each model per-cell using an unpaired student t-test. P-value thresholds are annotated as *: < .05, **: < .005, ***: < 0.0005, and ****: < 0.00005. The standard presence threshold of 0.5 is also noted as a dashed line. *deepbiosphere* and *MaxEnt* are the only two models to correctly predict *Q. lobata* as present and significantly higher

within the pixels annotated as such. The *Inception* baseline predicts no pixels as present, while the human annotators predicted non-QL pixels with a significantly higher cover annotation than QL pixels, which is problematic since all example images used to train the annotators were of *Q. lobata*. (B) The same analysis as in (A), but with *Q. agrifolia* and the corresponding Coast Live Oak alliance (QA). Human annotations do properly distinguish QA pixels from non-QA pixels, while *MaxEnt* and *Inception* fail to distinguish them readily. (C) The same analysis as in (A), but with *C. cuneatus* and the corresponding Ceanothus Chapparal alliance (CC). Again *deepbiosphere* and *MaxEnt* are the only two models to successfully disambiguate the alliance and a present threshold.

Fig. S23 | San Francisco local biodiversity case study.
(A) unobstructed NAIP imagery of the peninsula. (B) registered parks in San Mateo, San Francisco, and Marin counties (orange polygons). Parks data was sourced from the California state parks service, San Mateo County GIS warehouse, Marin GIS warehouse and San Francisco GIS warehouse.

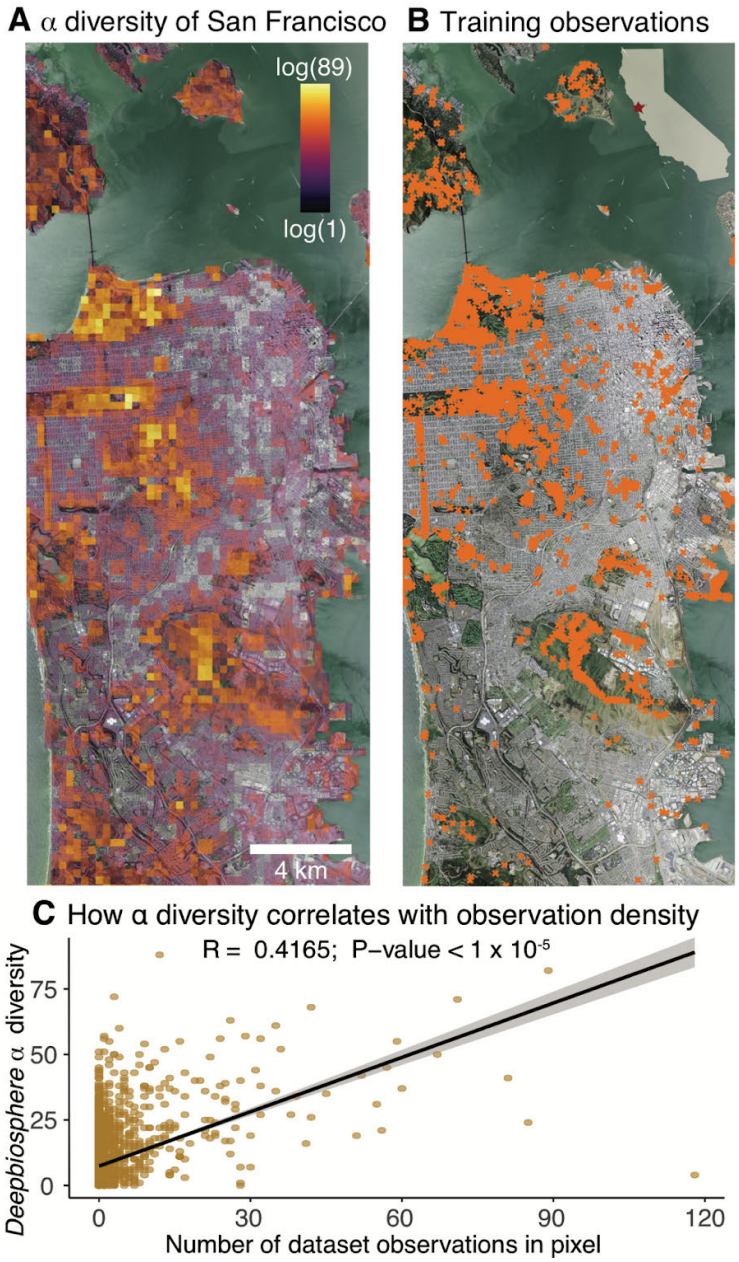

Fig. S24 | *deepbiosphere*$_{alpha}$ local biodiversity across San Francisco.
(A) Predicted alpha diversity from *deepbiosphere*. Values are derived by counting the number of species with a predicted presence of >= 50% then adding one and taking the logarithm of this value, to smooth out the color gradient for visualization. A transparency mask is also applied to the intensity at each pixel, with pixel transparency mapping predicted alpha diversity to 0-1. Finally, values are clipped to only include areas within the extent of the terrestrial California state boundary as defined in GADM version 4.0. (B) The distribution of observations from the training dataset across San Francisco. Observations from both the train and test set are noted in orange. (C) Correlation of predicted alpha diversity and number of dataset observations in a given 256 × 256m pixel. Number of observations is expected to correlate both with alpha diversity and park locations, since users can usually only upload observations on iNaturalist from publicly accessible locations or occasionally from private areas where they have access.

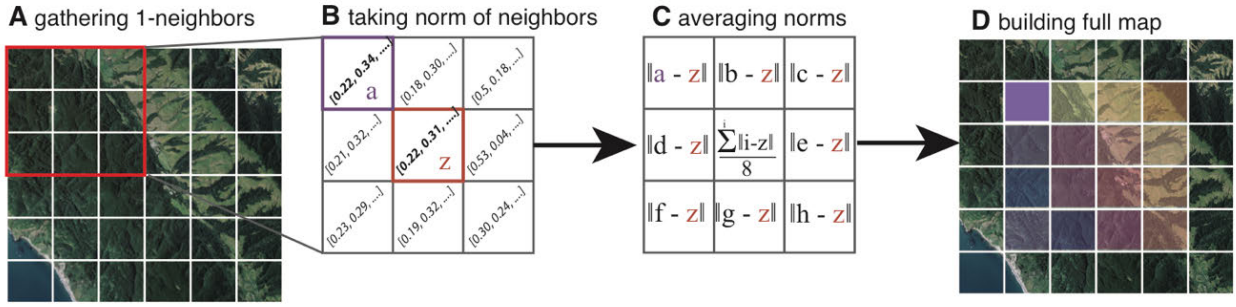

Fig. S25 | Beta diversity algorithm visual explanation.
Visual explanation of convolutional ecosystem transition detection algorithm inspired by edge detection filters in computer vision. Specifically, (A) we take the sigmoid-transformed probabilities from *deepbiosphere* and (B) for each pixel, take the difference between each neighboring cell and the central pixel, then (C) take the norm of these differences to generate the turnover for that pixel. Convolving this norm-of-neighbors pixel-by-pixel generates a 256m-resolution map of species turnover.

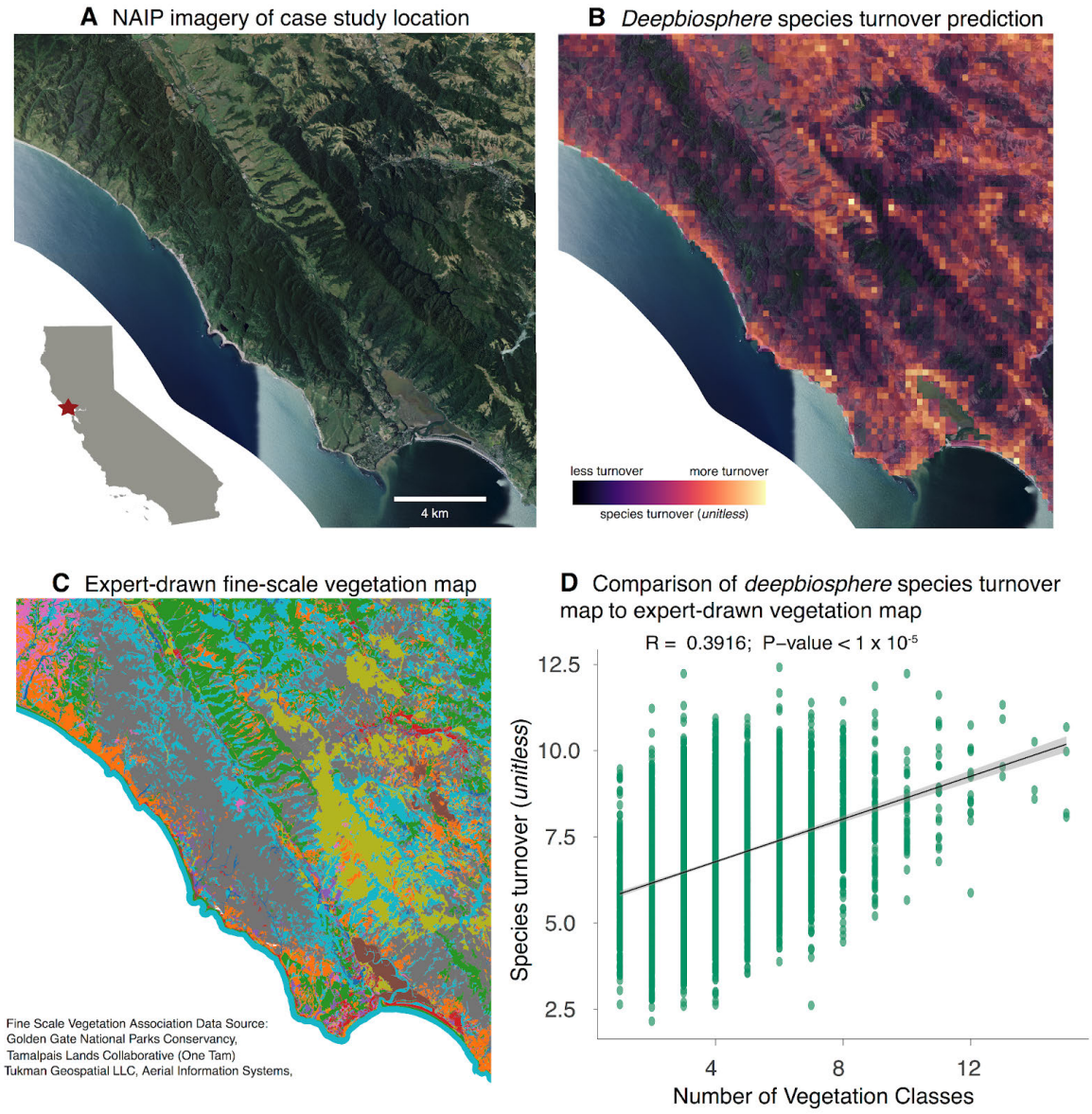

Fig. S26 | *deepbiosphere*$_{alpha}$ prediction of species turnover (beta diversity) in northern Marin county.

(A) This region of Marin county is located at the boundary between the Coast Range and Central California Foothills and Coastal Mountains ecoregions, thus making it an ideal location to study ecosystem transitions. (B) Aerial imagery of the area confirms visually that the landscape is highly varied with beach, forest, and grassland ecosystems visible. (C) Using the edge detection algorithm from **Fig. S25**, regions of high species turnover can be visually located. Specifically, pixels corresponding to areas along the beach and between the grassland and forest zones have a higher intensity, which conforms to intuition that boundaries between communities have high species turnover. (D) Mapping the vegetation classification of the area down to the association level confirms that this area indeed contains many unique ecosystems. The predicted ecosystem turnover shows a significantly strong correlation with the number of vegetation classes for a given pixel.

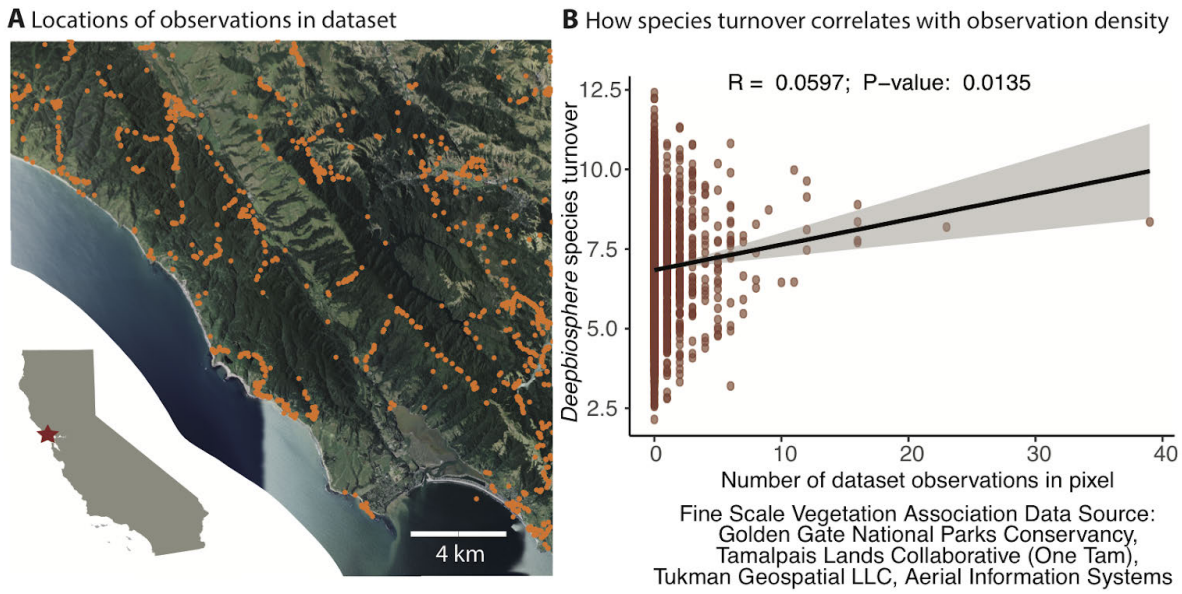

Fig. S27 | Relationship of beta diversity to observation density

Comparison of observation density to predicted beta diversity (also referred to as species turnover). Unlike alpha diversity, beta diversity predicted by $deepbiosphere_{alpha}$ does not correlate with the number of observations in a given area, as is to be expected. Encouragingly, this also shows that the beta diversity predicted by $deepbiosphere_{alpha}$ is not simply an artifact of the underlying data's spatial distribution, but more likely reflecting real change detection.

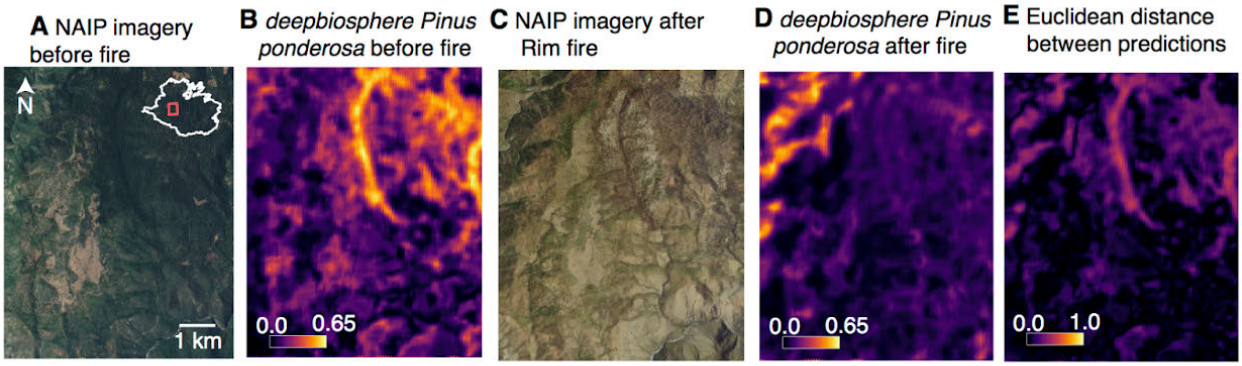

Fig. S28 | Example of change in species presence before and after Rim Fire

Example change in *deepbiosphere*$_{alpha}$ predicted presence for *Pinus ponderosa* before and after the Rim Fire. (A) NAIP imagery before fire in a subset of the fire's extent (location inside perimeter of fire inset). (B) Predicted probabilities for ponderosa pine (*Pinus ponderosa*) before the fire, using the imagery in (A). (C) NAIP imagery after the fire from the same geographic location. (D) Predicted probabilities for ponderosas after the fire. It's clear to see how the predicted presence of ponderosas decreases dramatically in most locations after the fire.(E) To capture this temporal trend, via the community change metric, we take the Euclidean distance between the probabilities in (B) and (D) per-pixel, doing so across all species in the dataset to generate a metric of change on the community level.

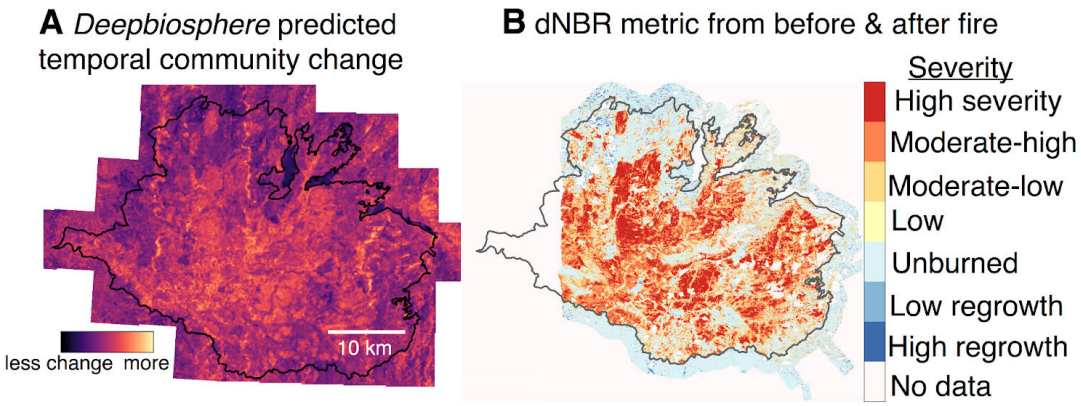

Fig. S29 | Comparison of *deepbiosphere*$_{alpha}$ community change metric to empirical nDBR metric. (A) Here we plot the temporal community change metric across the entire extent of the fire (perimeter in black). The community change metric is unitless but a brighter color implies a higher amount of change, while a darker implies less. The resolution per-pixel is 35 m per-side. Pixel intensity is derived by taking the distance between the predicted species' presence in 2012 and from 2014, before and after the fire. (B) Here we have plotted the difference in normalized burn ratio taken from MASTER hyperspectral data collected before the fire in 2013 and after in June of 2014. Higher dNBR implies that less photosynthesis is occurring and more char remains on the ground. Visually, dNBR and temporal community change appear to be highly correlated, which is intuitive as we expect areas with more severe burning to have experienced a more profound community change, as a larger percentage of mature trees are likely gone, and soil potentially may even be sterilized due to the intensity of the blaze, confounding species' ability to regrow post-blaze.

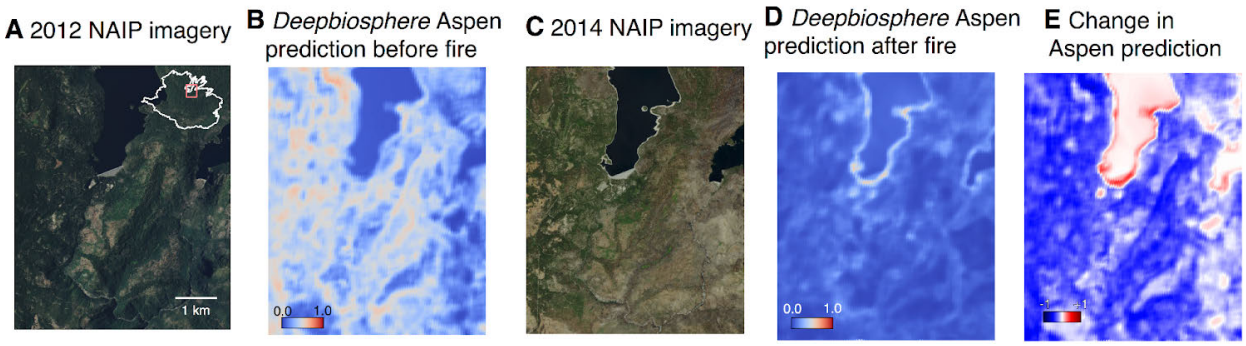

Fig. S30 | Decrease in predicted Aspen (*Populus tremuloides*) presence before and after the Rim Fire.

(A) NAIP imagery before the fire; (B) *deepbiosphere*$_{alpha}$ predictions of Aspen presence before the fire, using the imagery from (A). Blue colors represent expected absence with a threshold of 0.5, reddish colors indicate predicted presence. (C) NAIP imagery from after the fire. (D) *deepbiosphere*$_{alpha}$ predictions of Aspen presence after the fire, using the same scale as (B). (E) predicted change in Aspen presence (calculated by taking the difference between images (D) and (B)). More blue coloration implies a decrease in predicted presence while red implies an increase. Aspens as a species are well-adapted to fire-prone ecosystems and benefit from canopy-clearing fires. However, Aspen's strategy for fire survival is not that individual trees survive the blaze, but rather that the root system survives while parent trees burn, then sending up suckers into the clearing made in the canopy after the fire to dominate a local area's community until slower-growing conifer and pine species eventually outcompete the Aspen. Since Aspens are large trees and directly visible from aerial imagery, *deepbiosphere*$_{alpha}$ has likely learned to detect the crowns of adult Aspen directly, and thus the decrease in Aspen presence that *deepbiosphere*$_{alpha}$ predicts likely stems from the fact that *deepbiosphere*$_{alpha}$ can no longer detect the now-dead adult Aspen's canopy and the new suckers being sent up by the Aspen colony's root system aren't yet detectable from aerial imagery less than a year after the fire. As a note, there were only 90 examples of Aspen in *deepbiosphere*'s training data yet still predicts Aspen presence before the fire, a positive result given how little training data was provided.

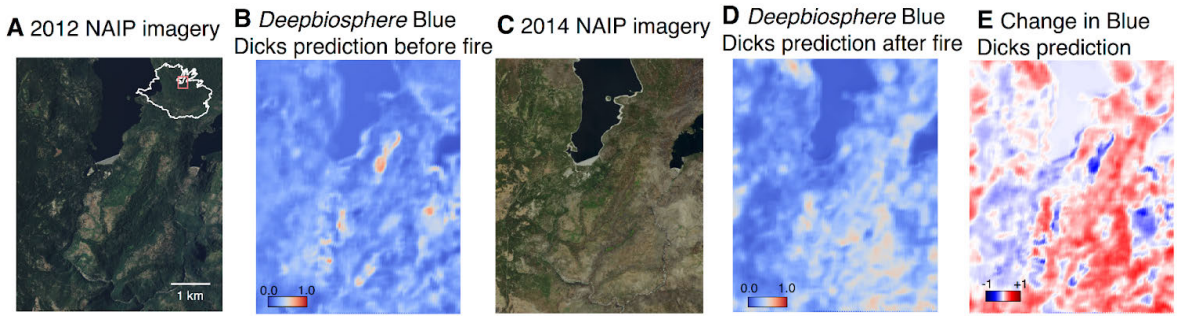

Fig. S31 | Increase in predicted a well-known fire follower perennial herb Blue Dicks (*Dichelostemma capitatum*) presence before and after the Rim Fire.

(A) NAIP imagery before the fire; (B) *deepbiosphere*$_{alpha}$ predictions of Blue Dicks presence before the fire, using the imagery from (A). Blue colors represent expected absence with a threshold of 0.5, reddish colors indicate predicted presence. (C) NAIP imagery from after the fire. (D) *deepbiosphere*$_{alpha}$ predictions of Blue Dicks presence after the fire, using the same scale as (B). (E) predicted change in Blue Dicks presence (calculated by taking the difference between images (D) and (B)). Blue Dicks are a small annual herbaceous flower typically found in grassy areas and are a well known "fire follower" species, often one of the first to grow and flower after a fire event. Although they are more of a chaparral species, their range does extend into the foothills of the Sierra and may in fact be the plant Hetch Hetchey valley was named after. *deepbiosphere*$_{alpha}$ predicts a marked increase of Blue Dicks presence after the fire, which is to be expected as the forest canopy is cleared by fire, allowing this small grass full sun access to grow. All in all, there are 3,731 training observations for Blue Dicks in the dataset, but only 37 observations are from the Sierra Mountains ecoregion, making it all the more impressive that *deepbiosphere*$_{alpha}$ accurately predicts this fire-follower's presence post-Rim fire.

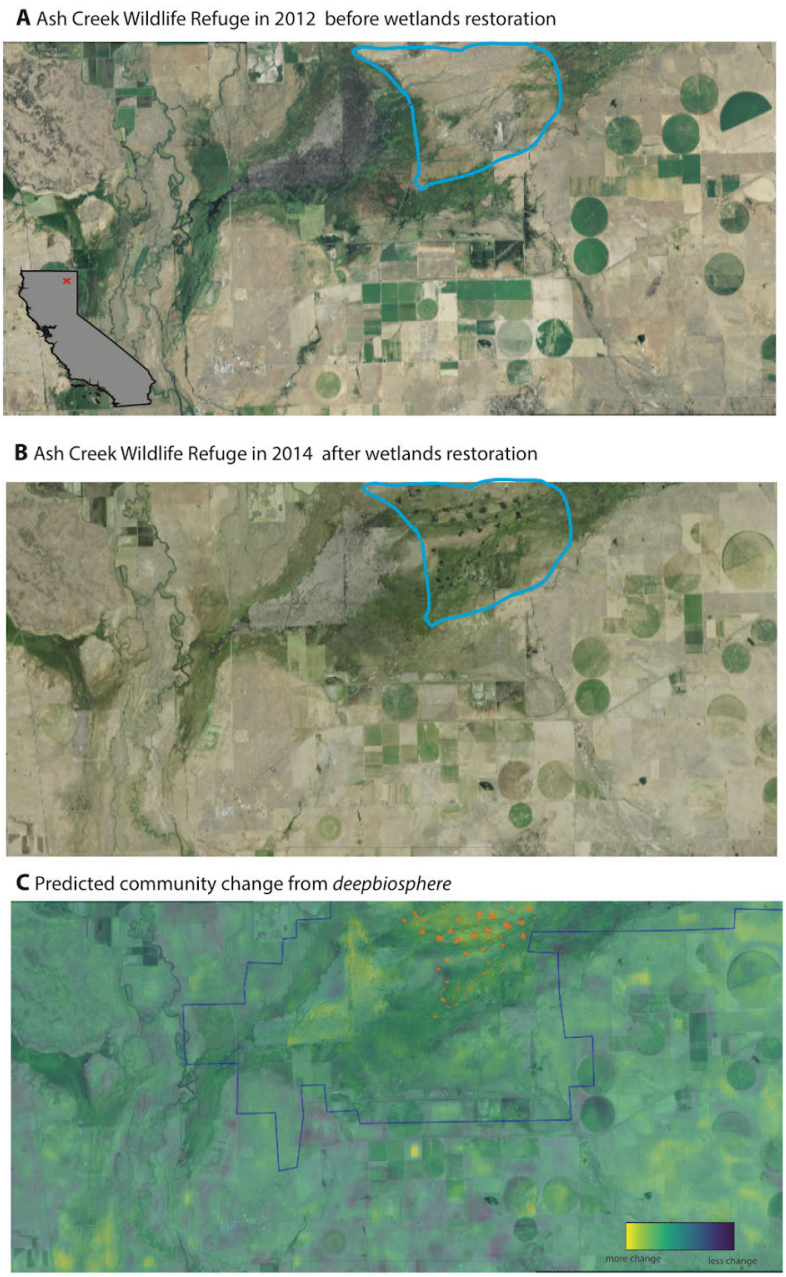

Fig. S32 | Detecting ecosystem change from wetlands restoration project using *deepbiosphere*$_{alpha}$. (A) Ash Creek Wildlife Refuge is located in northeastern California. Due to water diversion and channel cutting during the 20th century, this freshwater wetland was degraded, resulting in a drying of the wetlands and grassland incursion (highlighted in cyan). Park boundaries were acquired from the California Department of Fish & Wildlife. (B) In 2012, the refuge underwent a wetlands restoration project using a pond and plug strategy, which created 80 new ponds (orange polygons) and plugged the cut channel, resulting in the return of seasonal flooding that is critical for the health of marshy plants. This return of the native wetland system is clearly visible in (B), along with the ponds used to bring up the water table. (C) *Deepbiosphere's* community change predictions (unitless) along with the boundary of the park and the added ponds (ponds are marked in orange).

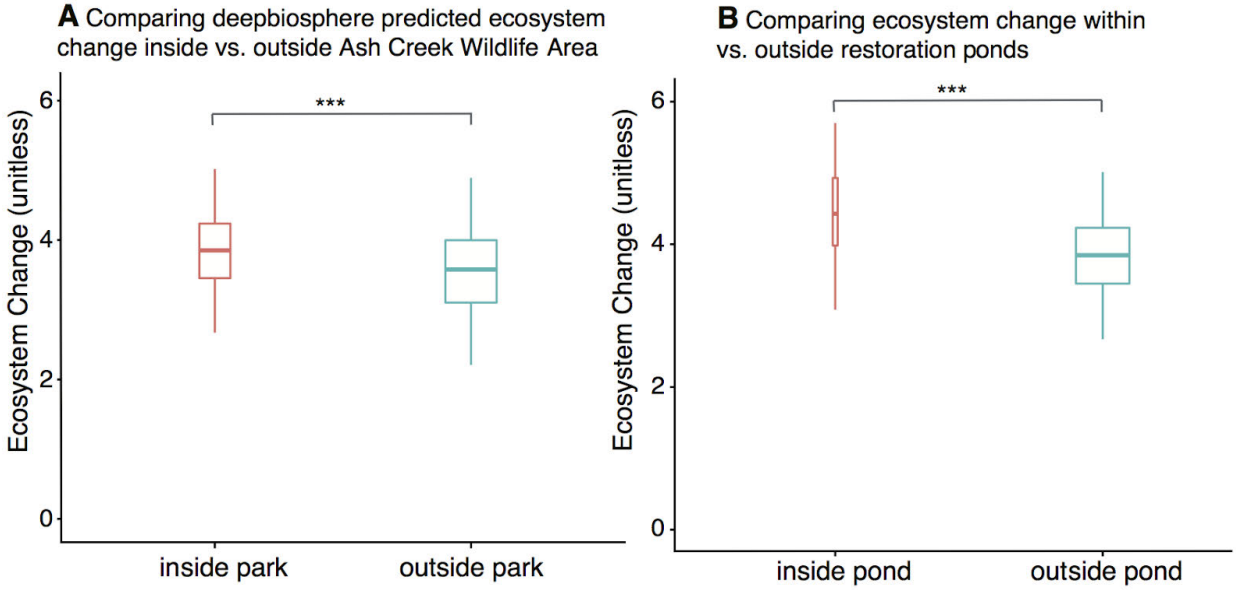

Fig. S33 | Comparing *deepbiosphere$_{alpha}$*'s ecosystem change predictions at Ash Creek National Wildlife Refuge.

(A) By comparing the pixel-wise values of deepbiosphere's ecosystem change predictions from inside the wildlife refuge to the values outside (blue boundary from **Fig. S33C**) we can see that while there is significant variation, the values on average are significantly higher inside of the park versus outside. This suggests that *deepbiosphere$_{alpha}$* is indeed detecting the deliberate ecosystem change initiated by the wetlands restoration project. P−values are unpaired t-test. (B) Comparing *deepbiosphere$_{alpha}$* ecosystem change predictions within added ponds. We can further compare the pixel-wise values of *deepbiosphere$_{alpha}$* ecosystem change predictions from inside the restoration ponds (orange polygons from **Fig. S33C**) to the same values from inside the rest of the park. The ecosystem change metric is significantly higher within the bounds of the ponds, indicating that *deepbiosphere$_{alpha}$* is indeed detecting real ecological change, as the areas within the ponds are known to have been deliberately transitioned from grasslands to marshy ponds as part of the wetlands restoration project.