PNAS COMMENTARY

Deep learning sharpens vistas on biodiversity mapping

Thomas J. Givnish[a,1]

Fig. 1. *Coreopsis gigantea* (Asteraceae) is a summer-deciduous shrub with thick, succulent stems; it is native to cool, dry, often foggy coastal dunes, hillsides, and bluffs along the Pacific coast from San Francisco to northwestern Baja California. The new deep-learning program *Deepbiosphere*—based partly on remote sensing data, which can easily track *Coreopsis*' wind-disturbed, low-coverage habitat—is 11 times as accurate in predicting the distribution of this species as *Maxent* based solely on climatic data. Image Credit: William T. Reid (photographer).

Deep learning—the use of artificial neural networks to detect and analyze patterns in complex data—is revolutionizing the use of computers for image and speech recognition, autonomous driving, financial projections, DNA sequencing, medical diagnosis and treatment, behavioral analyses, and the study and modeling of ecological processes (1, 2). One of the great strengths of deep learning—and its key to successful application across such a wide range of fields and problems—is that it specifies no model in advance and pursues whatever approach best connects potential predictors in input data to desired predictions, with appropriate measures taken to avoid overfitting.

In ecology, remote sensing is creating a second revolution, allowing the massively parallel collection of spatially structured data at a variety of scales on species identity, functional traits, chemistry, physiology, phylogenetic affinity, species interactions, community composition, structure, diversity, and ecosystem productivity, as well as patterns of change through time in each of these attributes (3–5). One of the major challenges facing ecological studies that employ remote sensing, however, has been tracking rare or understory species that contribute few if any reflected or emitted photons that can be detected by sensors on drones, aircraft, or satellites.

Author affiliations: [a]Department of Botany, University of Wisconsin-Madison, Madison, WI 53706

Author contributions: T.J.G. wrote the paper.

The author declares no competing interest.

See companion article, "Deep learning models map rapid plant species changes from citizen science and remote sensing data," 10.1073/pnas.2318296121.

[1]Email: givnish@wisc.edu.

Published September 30, 2024.

Gillespie et al. (6) provide a novel and highly powerful solution to this challenge. They use deep learning to predict occurrences of thousands of common, rare, or understory plant species at scales from meters to hundreds of kilometers, based on remote sensing data (in this first case, high-resolution aerial photographs shot in color and infrared), climatic data, and hundreds of thousands of geolocated species occurrences drawn from herbarium records and (especially) from citizen-science observations logged on the widely used database iNaturalist. The authors tested this approach mainly in California, which is an especially apt venue given its high density of species, species observations, and remote sensing data (Fig. 1).

Their technique builds on a few recent papers that have also applied deep learning or other forms of AI to map species occurrences. However, Gillespie et al. substantially improve the power and range of analyses by relating species occurrences to both climatic *and* remotely sensed data, by training their networks with the distributions of species *and* related taxa at the genus and family levels, and by including information on the distributions of co-occurring species. This approach takes advantage of two key findings in community ecology: Close relatives tend to have similar ecologies and distributions, and joint species occurrences refine estimates of the positions of species and samples along environmental gradients. In addition, using both climatic and remotely sensed data allows species presence and community composition to be predicted accurately at a wide range of spatial scales, reflecting variation in temperature, rainfall, and seasonality over tens to hundreds of kilometers of distance and hundreds to thousands of meters in elevation, while differences in reflectance or emission at different wavelengths can track variation in disturbance, succession, substrate, and land use—and consequent differences in species occurrence—at very fine scales, down to a few meters.

The accuracy and speed of current and future versions of *Deepbiosphere* could contribute to the massive international efforts that will be needed to monitor biodiversity and ecological services on a global scale under the Convention on Biological Diversity.

Gillespie et al. provide several compelling applications of this approach, using their *Deepbiosphere* model of convolutional neural networks to predict the distributions of coast redwoods in northern California, various trees and shrubs in the Mediterranean climate region in southern California, understory plants of young and old-growth redwood stands, and vegetation and abrupt ecotones in Marin County, together with estimates of the tempo of community change across larges areas associated with the severe Rim Fire in the western Sierra Nevada foothills in 2013. Predicted differences in the distributions of Pacific blackberry and redwood sorrel in young vs. old-growth redwood forests are highly significant, and track changes in forest age and structure are evident over just a few meters in the high-resolution RBG+IR aerial images used. Across 34 species examined statewide, *Deepbiosphere* provides much greater resolution of distributions at subkilometer scales than other approaches. Furthermore, *Deepbiosphere* is often more accurate than other models often used to predict species distributions. For example, based on the area-under-the-receiver-curve metric, *Deepbiosphere* had a minimum AUC_{ROC} of 0.986, compared to 0.04 for *Maxent*. Based on this comparison, *Deepbiosphere* has a 9.7% ± 26.5% advantage. However, most of that advantage is due to far poorer performance by *Maxent* on four species: *Lupinus arboreus*, *Malacothrix saxatilis*, *Coreopsis gigantea*, and *Rhus integrifolia*. All are associated with low-coverage habitats in coastal dunes, coastal sage, chaparral, and cliffs, readily recognized by *Deepbiosphere* but missed by *Maxent* and other widely used species distribution models (SDMs).

Across 11 different metrics, *Deepbiosphere* usually adds accuracy to distributions over those predicted by other SDMs, including *Maxent*, *Inception*, and *Random Forest*. *Deepbiosphere* shows the most consistent performance across species, especially for rare species. *Deepbiosphere* based on both climatic and remote sensing data outperforms versions on only climatic data or only remote sensing data. Furthermore, versions that include phylogenetic and associational data—that is, training based on related species in the same genus or family, and on associated species—outperform others.

The Gillespie et al. approach, based on widely available, relatively low-cost data, can be pursued over much larger areas than California, use much larger fractions of the Tree of Life, and track dynamics based on remote sensing data gathered at decadal frequencies. *Deepbiosphere* is a powerful analytical pipeline that can easily be improved in the future. Exciting improvements might include 1) use of more informative multi- or hyperspectral remotely sensed data; 2) inclusion of soil or bedrock data, given the importance of edaphic conditions in limiting the distributions of many species; and especially 3) incorporating modules that use higher-level spatial and temporal data—especially on patch size, landscape context, distance from similar habitats, dispersal corridors, species interactions, and time since disturbance—to improve predicted distributions of species strongly affected by fragmentation, edge effects, and successional dynamics. Hyperspectral data, although expensive to obtain, offers big advantages in species identification and quantification of leaf chemistry, vegetation structure, and community composition and diversity (3–5). Gillespie et al. themselves show that including substrate data substantially increases the accuracy of predicted distribution for some species. Finally, many ecologists and conservation biologists have used experimental or observational data to show that fragmentation, landscape context, and species interactions can have important impacts on species distributions. For example, Terborgh et al. (7) demonstrated how isolation can cause plant species loss from forest fragments via loss of mega- and mesopredators. Rogers et al. (8) showed that landscape context and fragmentation have increasingly accounted for forest composition in Wisconsin—and local environmental conditions increasingly less—in moving from the 1950s to 2000s. Damschen et al. (9) demonstrated how corridors elevate plant species richness in patchy landscapes. All these phenomena might be

captured in upcoming editions of *Deepbiosphere*. More broadly, the accuracy and speed of current and future versions of *Deepbiosphere* could contribute to the massive international efforts that will be needed to monitor biodiversity and ecological services on a global scale under the Convention on Biological Diversity (4, 10).

1. M. L. Borowiec *et al.*, Deep learning as a tool in ecology and evolution. *Methods Ecol. Evol.* **13**, 1640–1666 (2022).
2. J. Egger *et al.*, Medical deep learning–A systematic meta-review. *Comput. Methods Biomed.* **221**, 106874 (2022).
3. G. P. Asner *et al.*, Quantifying forest canopy traits: Imaging spectroscopy versus field survey. *Remote Sens. Environ.* **158**, 15–27 (2015).
4. J. Cavender-Bares *et al.*, Integrating remote sensing with ecology and evolution to advance biodiversity conservation. *Nat. Ecol. Evol.* **6**, 506–519 (2022).
5. R. R. Carlson *et al.*, The effect of reef morphology on coral recruitment at multiple spatial scales. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2311661121 (2024).
6. L. E. Gillespie, M. Ruffley, M. Expósito-Alonso, Deep learning models map rapid plant species changes from citizen science and remote sensing data. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2318296121 (2024).
7. J. Terborgh *et al.*, Ecological meltdown in predator-free forest fragments. *Science* **294**, 1923–1926 (2001).
8. D. A. Rogers *et al.*, Paying the extinction debt in southern Wisconsin forest understories. *Conserv. Biol.* **23**, 1497–1506 (2008).
9. E. I. Damschen *et al.*, Ongoing accumulation of plant diversity through habitat connectivity in an 18-year experiment. *Science* **365**, 1478–1480 (2019).
10. S. Díaz *et al.*, Set ambitious goals for biodiversity and sustainability. *Science* **370**, 411 (2020).