Scaling Up Agricultural Research With Artificial Intelligence

Brandon T. Bestelmeyer
USDA-ARS Jornada Experimental Range

Guillermo Marcillo
USDA-ARS

Sarah E. McCord
USDA-ARS

Steven Mirsky
USDA-ARS

Glenn Moglen
USDA-ARS

Lisa G. Neven
USDA-ARS

Debra Peters
USDA-ARS

Clement Sohoulande
USDA-ARS

Tewodros Wakie
USDA-ARS

***Abstract*—Agricultural systems are enormously variable in space and time. New and developing artificial intelligence (AI)-based tools can leverage site-based science and big data to help farmers and land managers make site-specific decisions. These tools are improving information about soils and vegetation that forms the basis for investments in management actions, provides early warning of pest and disease outbreaks, and facilitates the selection of sustainable cropland management practices. Continued progress with AI will require more observational data across a wide range of agricultural settings, over long time periods.**

Scaling up the results of site-based research to improve the efficiency and sustainability of agricultural systems at national to global scales is a primary scientific challenge. Some agricultural innovations, such as new crop cultivars, are so uniformly beneficial that the rate of spread in their adoption is limited largely by the transfer of information. Agricultural "technology transfer" and collaborative science are used to convey information to producers and land managers and to assist them in adapting innovations to regional production systems.[1] In some cases, however, the

Digital Object Identifier 10.1109/MITP.2020.2986062
Date of current version 21 May 2020.

Published by the IEEE Computer Society

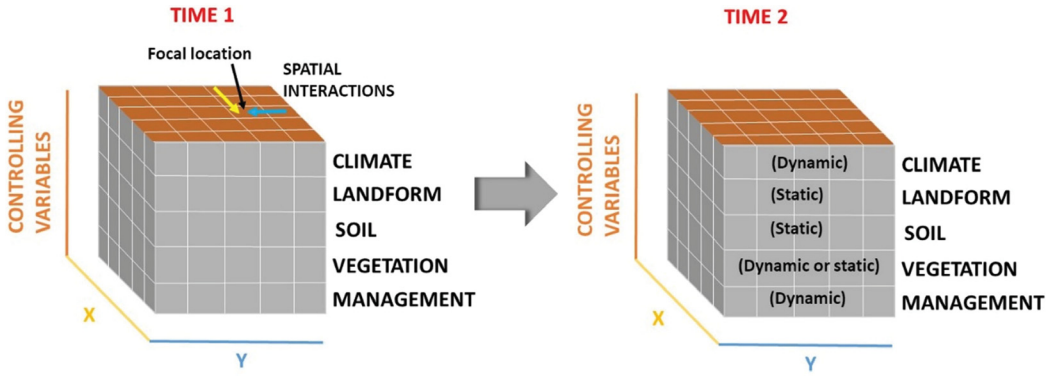

Figure 1. Schematic of data structure used in broad-scale agricultural AI models. Grid cells or point locations in *x*–*y* space (which may influence one another via spatial interactions) have multiple layers of potential covariates (controlling variables), some of which may change value from one time period to the next.

scientific information provided to managers is highly sensitive to varying spatial and temporal contexts. In such cases, information products should be tailored to specific locations and times by linking a body of site-based observations to computational models. Such models can fill information gaps between sparse local observations using gridded datasets representing key controlling variables, such as climate, soils, vegetation, land use, hydrology, neighborhood effects (spatial interactions), and other variables (see Figure 1).

The potential utility of AI-powered models to scale or extrapolate information has increased due to new AI/machine learning algorithms and the availability of "big" spatiotemporal datasets representing many variables.[2] In the following, we highlight recent examples in which AI modeling approaches and related tools can provide precision information to agricultural producers and land managers across broad spatial extents. We also highlight future directions and the implications of AI for site-based research within networks, such as the system of U.S. Department of Agriculture (USDA) Agricultural Research Service research stations and the USDA Long-Term Agroecosystem Research (LTAR) network.

PRECISION ENVIRONMENTAL INFORMATION

Spatially explicit information about soil classes, properties, and vegetation is the basis for decisions about private land management decisions by farmers and ranchers, public land management by the U.S. government, and governmental conservation policy support. All told, the cost of these decisions measures in the billions of dollars per year. Precise information about soils and vegetation to guide these decisions, however, has been challenging to obtain. Conventional soil mapping, especially in the extensive rangelands of the western United States, often provides only a coarse estimate of soil properties at a specific location. Many parts of the world lack any soil maps. Similarly, key attributes of vegetation, such as the composition and productivity of plant species, are only coarsely mapped or not mapped at all. To provide more precise estimates of soil properties, machine learning algorithms have incorporated data from over one hundred thousand field soil samples gathered across the world. The algorithms have also incorporated covariates including gridded remotely sensed and modeled data on factors that control soil formation, such as climate, landform, hydrology, and vegetation cover. A global product, SoilGrids, provides estimates of nine soil properties at standard depths at a 250-m resolution.[3] These products also represent the uncertainty of soil predictions. Products of global extent may have higher uncertainty than nationally, or regionally, tailored models that can capitalize on regional covariates and local knowledge of soil property–covariate relationships. Thus, "digital soil mapping" is being performed and refined at subglobal and regional extents and at finer resolutions.[4]

A similar approach is being applied to map the dynamic composition of vegetation across rangelands in the western United States at a

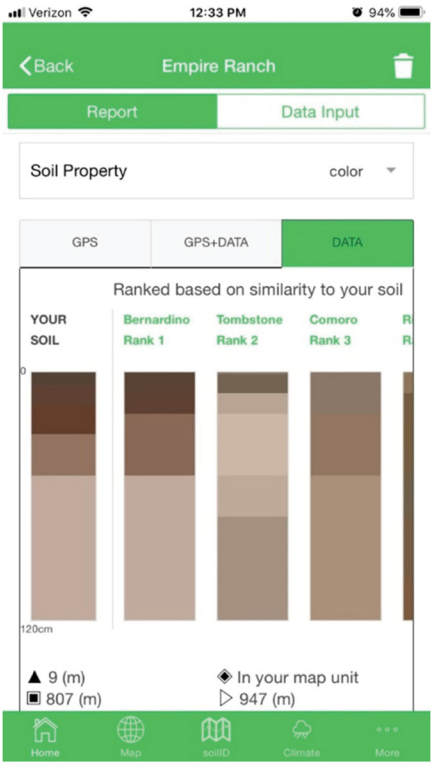

Figure 2. Example of soil class prediction delivered to mobile application (LandPKS) via AI-powered cloud computing (based on[7]).

30-m resolution, using tens of thousands of standardized local observations of vegetation and remotely sensed and modeled covariates.[5] Using the computational power of Google Earth Engine, Landsat imagery from 1984 to 2017 constitutes the basis for yearly predictions of vegetation cover, which users can query and visualize with a custom web application (https://rangelands.app/).

Continuous soil and vegetation predictions, in turn, can be combined with other models to predict and scale up processes, such as soil erosion. For example, bare soil cover, canopy gap distribution, and vegetation height estimates modeled in fractional cover products can be used as inputs in a sediment transport model to produce spatially explicit dust flux estimates.[6]

Vegetation and soil classification accuracy is ultimately limited by the availability of training data and the utility of available spatially continuous covariates, but new mobile applications can aid users in collecting local data that can be integrated with machine learning models. For example, the USDA-ARS unit in Las Cruces, NM developed a mobile application (the Land Potential Knowledge System; LandPKS) that guides users in collecting data on several soil properties at a location.[7] Global positioning system (GPS) locations provided by a cell phone are used to query soil sample and soil covariate databases to produce a local soil database (see Figure 2). The soil data the user enters and covariates mapped at the user's location are incorporated into a local automated machine learning model (AutoML) to predict the most probable soil class at the user's location. Precision environmental information will increasingly combine user inputs with existing observations and information from covariate databases.

PEST AND DISEASE PREDICTION AT REGIONAL TO NATIONAL SCALES

One of the most important challenges in agriculture is managing the impact of invasive pests and pathogens. As transportation allows people and products to move, it also provides a pathway for pests and pathogens to expand their range. AI and machine learning can assist in identifying those areas most at risk of invasions/outbreaks as well as assisting in plans to mitigate the spread of invasives or diseases. Ecological niche modeling (ENM) has greatly expanded with the use of machine learning and availability of gridded covariate data. Machine learning based models, such as MaxEnt,[8] can be used to identify suitable habitats where species could establish and reproduce.

USDA-ARS scientists in Wapato, WA, have been using ENM to identify areas in the United States and internationally where either currently established pests could expand their range in response to climate change or newly arriving invasive species could potentially spread. These maps can assist in decision making for international trade policies that recognize a pest of concern, but in which the ENM shows little or no suitable habitat to support this pest. This information can be invaluable during trade negotiations with prospective importing countries.

Such machine learning based models have proven their utility in real time. For example, the range of the oriental fruit moth (OFM), *Grapholita molesta*, in Washington State was predicted

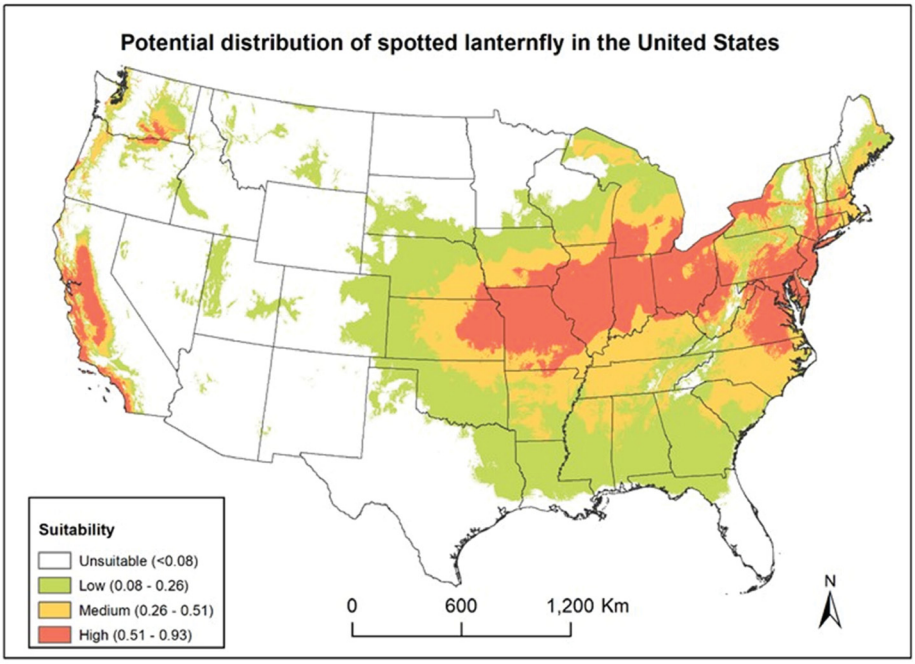

Figure 3. Predicted distribution of *L. delicatula* in the United States from a machine learning model. Areas shaded in red, yellow, and green indicate high, medium, and low habitat suitability, respectively. White areas are unsuitable for *L. delicatula* establishment. See.[10]

to be increasing due to climate change.[9] At the end of the 2018 crop year, multiple reports of OFM damage began pouring in from growers and pest management companies. The models were developed using a conservative climate change scenario; the actual expansion of this pest's range occurred about five years earlier than the model predicted.

A similar ENM model identified areas in the United States with suitable habitat to support a newly invasive species, the spotted lanternfly, *Lycorma delicatula*[10] (see Figure 3). Following the publication of this model, new populations of spotted lanternfly were discovered in areas that the model predicted as highly suitable habitat. These examples highlight the utility of AI for predicting invasive species spread with sufficient lead time to develop mitigation plans.

Animal disease modeling has also benefitted from machine learning. A transdisciplinary team of USDA scientists evaluated a large suite of spatially distributed environmental covariates (>400) using MaxEnt to develop early warning strategies for vesicular stomatitis (VS), a common viral vector-borne vesicular disease affecting livestock throughout the Americas.[11] VS occurrence at the scale of individual landowners was related to conditions that can be monitored (i.e., rainfall, temperatures, and streamflow) or modified (i.e., vegetation). On-site green vegetation during the month of occurrence and higher rainfall four months prior combined with either cool daytime (disease expansion) or nighttime (disease incursion) temperatures one month prior were common predictors of VS occurrence. At landscape to regional scales, conditions that favor specific VS biological vectors were predicted, including black flies in incursion years and biting midges in expansion years.

LOCATION-SPECIFIC FORECASTS OF CROP PERFORMANCE

A primary goal of sustainable food production systems research is to devise strategies to increase agricultural outputs while improving soil health, water quality, pest resistance, and resilience to climate change.

The use of cover crops is one of the most important sustainability strategies in crop production systems. Cover crops are nonmarketable plants that grow in between cash crop

plantings and can increase crop productivity while providing other environmental benefits to the farmer and the public. Realizing these benefits, however, is largely dependent on good cover crop performance. Farmers need sound early-season information to reduce the risk of cover crop failure and to offset cover crop planting costs by maximizing the benefits accrued through their use.

Current recommendation systems for cover crop adoption are based on expert opinions, or are linked to agronomic simulation models. Common agronomic tools, such as process-based models that capture soil and crop responses in detail, have been successfully applied to some types of crops. Unfortunately, process-based models adapted to simulate cover crops have shown only limited ability to predict growth and development. For example, a process-based model accurately predicted biomass of a typical small grain cover crop (cereal rye, *Secale cereale*), in only five out of ten years when compared with field observations.[12] AI can improve predictions by taking advantage of large datasets from real-time sources (i.e., remote sensors, digital farm equipment, and satellites). Better predictions can lead to more widespread adoption of cover crops and expansion of the benefits they provide.

USDA-ARS scientists in Beltsville, MD, are conducting cutting-edge research to predict the spatial and temporal variation of cover crops and their effects on crop systems. For example, remote sensing and crop management and performance data from a cereal rye cover crop were compiled from three years of Maryland and Pennsylvania field experiments testing rye response to nitrogen fertilizer. Using this dataset, a machine learning model (Random Forest) was trained and optimized to predict biomass of the cover crop. Testing of the model using validation data from a study in North Carolina revealed that 60% of biomass predictions corresponded to ranges of observed ground-truth biomass, reaching accuracy levels that surpass those reported via previous process-based modeling. Furthermore, crop modelers and data scientists in Beltsville are collating cover crop datasets from across the country, complementing an extensive on-farm soil nitrogen and water monitoring network featuring high-resolution imagery acquisition, to build the next generation of models. These new models will be used to create spatial maps of cover crop impacts on farm productivity for the first time.

CONCLUSION

Our review illustrates how AI-based tools can deliver a variety of high-quality, site-specific information products to producers and managers across broad spatial extents. The difficulty in increasing the practical utility of these tools reflects general challenges associated with AI-based technologies, but we want to highlight two key problems.

First, even though the availability of big data has made AI potentially useful, we often do not have enough data to provide predictions with desired accuracy given the high spatiotemporal variability inherent to agricultural systems at broad scales. We need more observations of the phenomena we seek to predict in order to train better models. These observations must be gathered across the breadth of spatial variation to which models are applied and should be long-term to account for dynamic controlling variables, such as climate, management, and lag effects. Research networks, such as LTAR, can contribute these observations from research sites, but such sites represent a limited range of variability. Observations from farms and ranches across landscapes and regions are needed, which can be facilitated by mobile technologies and collaborative networks involving farmers and ranchers.

Second, standardized methods and data integration, harmonization, and availability are essential to sustain the AI revolution. These datasets include high-quality observations of phenomena of interest (such as species occurrence or field estimates of plant production) as well as remotely sensed covariates, which can be processed in different ways. This is already a primary emphasis of the agricultural and ecological science communities, but the importance of bringing additional data into use, promoting a culture of data sharing among scientists, and providing systems to discover data and learn from repeated model-building processes[13] cannot be overemphasized.

ACKNOWLEDGMENTS

The authors would like to thank H. Savoy, P. Heilman, and M. Branstetter for providing early ideas in framing this article. This work was supported by appropriated funds to the USDA-ARS.

REFERENCES

1. J. B. Passioura, "Scaling up: The essence of effective agricultural research," *Functional Plant Biol.*, vol. 37, pp. 585–591, 2010.
2. D. P. C. Peters, K. M. Havstad, J. Cushing, C. Tweedie, O. Fuentes, and N. Villanueva-Rosales, "Harnessing the power of big data: Infusing the scientific method with machine learning to transform ecology," *Ecosphere*, vol. 5, 2014, Paper art67.
3. T. Hengl *et al.*, "SoilGrids250m: Global gridded soil information based on machine learning," *PLoS One*, vol. 12, 2017, Art. no. e0169748.
4. J. Maynard *et al.*, "Digital mapping of ecological land units using a nationally scalable modeling framework," *Soil Sci. Soc. Amer. J.*, vol. 83, pp. 666–686, 2019.
5. M. O. Jones *et al.*, "Innovation in rangeland monitoring: Annual, 30 m, plant functional type percent cover maps for U.S. Rangelands, 1984–2017," *Ecosphere*, vol. 9, 2018, Art. no. e02430.
6. N. P. Webb *et al.*, "Indicators and benchmarks for wind erosion monitoring, assessment and management," *Ecol. Indicators*, vol. 110, 2020, Art. no. 105881.
7. J. E. Herrick *et al.*, "Two new mobile apps for rangeland inventory and monitoring by landowners and land managers," *Rangelands*, vol. 39, pp. 46–55, 2017.
8. J. Elith, S. J. Phillips, T. Hastie, M. Dudík, Y. E. Chee, and C. J. Yates, "A statistical explanation of MaxEnt for ecologists," *Diversity Distrib.*, vol. 17, pp. 43–57, 2011.
9. L. G. Neven, S. Kumar, W. L. Yee, and T. Wakie, "Current and future potential risk of establishment of *Grapholita molesta* (Lepidoptera: Tortricidae) in Washington State," *Environ. Entomology*, vol. 47, pp. 448–456, 2018.
10. T. T. Wakie, L. G. Neven, W. L. Yee, and Z. Lu, "The establishment risk of *Lycorma delicatula* (Hemiptera: Fulgoridae) in the United States and globally," *J. Econ. Entomology*, vol. 113, pp. 306–314, 2019.
11. D. P. C. Peters *et al.*, "Developing multi-scale early warning strategies for vector-borne disease using big data-model integration and machine learning," *Ecosphere*, in press.
12. G. S. Marcillo, S. Carlson, M. Filbert, T. Kaspar, A. Plastina, and F. E. Miguez, "Maize system impacts of cover crop management decisions: A simulation analysis of rye biomass response to planting populations in Iowa, U.S.A," *Agricultural Syst.*, vol. 176, 2019, Art. no. 102651.
13. D. P. C. Peters *et al.*, "An integrated view of complex landscapes: A big data-model integration approach to transdisciplinary science," *Bioscience*, vol. 68, pp. 653–669, 2018.

Brandon Bestelmeyer is currently a Supervisory Research Ecologist with the Jornada Experimental Range, USDA-ARS, Las Cruces, NM, USA. Contact him at Brandon.Bestelmeyer@usda.gov.

Guillermo Marcillo is currently a Cropping System Modeler and a Data Scientist with the Beltsville Agricultural Research Center, USDA-ARS, Beltsville, MD, USA. Contact him at guillermo.marcillo@usda.gov.

Sarah McCord is currently a Computational Biologist with the Jornada Experimental Range, USDA-ARS, Las Cruces, NM, USA. Contact her at sarah.mccord@usda.gov.

Steven Mirsky is currently a Research Ecologist with Beltsville Agricultural Research Center, USDA ARS, Beltsville, MD, USA. Contact him at steven.mirsky@usda.gov.

Glenn Moglen is currently a Supervisory Research Hydrologist with the Beltsville Agricultural Research Center, USDA ARS, Beltsville, MD, USA. Contact him at glenn.moglen@usda.gov.

Lisa Neven is currently a Research Leader/Research Entomologist with the Temperate Tree Fruit and Vegetable Research Unit, USDA-ARS, Wapato, WA, USA. Contact her at lisa.neven@usda.gov.

Debra Peters is currently an Acting Chief Science information officer with the USDA ARS, Las Cruces, NM, USA, and a research scientist with Jornada Experimental Range, USDA-ARS. Contact her at deb.peters@usda.gov.

Clement Sohoulande is currently a Research Agricultural Engineer with the Coastal Plains Soil, Water, and Plant Research Center, Florence, SC, USA. Contact: clement.sohoulande@usda.gov.

Tewodros Wakie is currently a Research Ecologist with the Temperate Tree Fruit and Vegetable Research Unit, USDA-ARS, Wapato, WA, USA. Contact him at tewodros.wakie@usda.gov.