PERSPECTIVES

Perspectives on Artificial Intelligence for Predictions in Ecohydrology

ELIAS C. MASSOUD,[a] FORREST HOFFMAN,[a] ZHENG SHI,[b] JINYUN TANG,[c] ELIE ALHAJJAR,[d] MALLORY BARNES,[e] RENATO K. BRAGHIERE,[f,g] ZOE CARDON,[h] NATHAN COLLIER,[a] OCTAVIA CROMPTON,[i] P. JAMES DENNEDY-FRANK,[j] SAGAR GAUTAM,[k,l] MIQUEL A. GONZALEZ-MELER,[m] JULIA K. GREEN,[n] CHARLES KOVEN,[c] PAUL LEVINE,[f] NATASHA MACBEAN,[o] JIAFU MAO,[p] RICHARD TRAN MILLS,[q] UMAKANT MISHRA,[k,l] MARUTI MUDUNURU,[r] ALEXANDRE A. RENCHON,[s,g] SARAH SCOTT,[t] ERICA R. SIIRILA-WOODBURN,[j] MATTHIAS SPRENGER,[j] CHRISTINA TAGUE,[u] YAOPING WANG,[p] CHONGGANG XU,[v] AND CLAIRE ZARAKAS[w]

[a] *Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee*
[b] *Department of Microbiology and Plant Biology, University of Oklahoma, Norman, Oklahoma*
[c] *Climate Sciences Department, Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California*
[d] *Engineering and Applied Sciences Division, RAND Corporation, Arlington, Virginia*
[e] *O'Neill School of Public and Environmental Affairs, Indiana University Bloomington, Bloomington, Indiana*
[f] *Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California*
[g] *Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California*
[h] *Marine Biological Laboratory, Ecosystems Center, Woods Hole, Massachusetts*
[i] *USDA–ARS Hydrology and Remote Sensing Laboratory, Beltsville, Maryland*
[j] *Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, Berkeley, California*
[k] *Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Berkeley, California*
[l] *Bioscience Division, Sandia National Laboratories, Livermore, California*
[m] *Department of Biological Sciences and Earth and Environmental Sciences, University of Illinois at Chicago, Chicago, Illinois*
[n] *Department of Environmental Science, The University of Arizona, Tucson, Arizona*
[o] *Department of Geography, Indiana University Bloomington, Bloomington, Indiana*
[p] *Environmental Sciences Division and Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, Tennessee*
[q] *Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, Illinois*
[r] *Atmospheric Sciences and Global Change Division, Pacific Northwest National Laboratory, Richland, Washington*
[s] *Environmental Science Division, Argonne National Laboratory, Lemont, Illinois*
[t] *Thermal/Fluid Science and Engineering, Sandia National Laboratories, Livermore, California*
[u] *Bren School of Environmental Science and Management, University of California, Santa Barbara, California*
[v] *Earth and Environmental Sciences Division, Los Alamos National Laboratory, New Mexico*
[w] *Department of Atmospheric Sciences, University of Washington, Seattle, Washington*

(Manuscript received 26 January 2023, in final form 23 August 2023, accepted 25 August 2023)

ABSTRACT: In November 2021, the Artificial Intelligence for Earth System Predictability (AI4ESP) workshop was held, which involved hundreds of researchers from dozens of institutions. There were 17 sessions held at the workshop, including one on ecohydrology. The ecohydrology session included various breakout rooms that addressed specific topics, including 1) soils and belowground areas; 2) watersheds; 3) hydrology; 4) ecophysiology and plant hydraulics; 5) ecology; 6) extremes, disturbance and fire, and land-use and land-cover change; and 7) uncertainty quantification methods and techniques. In this paper, we investigate and report on the potential application of artificial intelligence and machine learning in ecohydrology, highlight outcomes of the ecohydrology session at the AI4ESP workshop, and provide visionary perspectives for future research in this area.

KEYWORDS: Ecology; Carbon cycle; Hydrologic cycle; Artificial intelligence; Machine learning

1. Introduction

Research in ecohydrology bridges the gap between ecosystem ecology and water cycle science by incorporating knowledge of land surface processes, plant physiology, atmospheric science, and hydrology (Rodriguez-Iturbe 2000; Asbjornsen et al. 2011; Guswa et al. 2020). Ecohydrology encompasses major components of the Earth system including vegetation, microbes, aquatic organisms, soils, the atmosphere, and surface and subsurface hydrology. Plant evapotranspiration and water use (Wang et al. 2019; Zhang et al. 2022), plant

Dennedy-Frank's current affiliation: Department of Marine and Environmental Sciences, Northeastern University, Boston, Massachusetts.

Corresponding author: Elias C. Massoud, massoudec@ornl.gov

DOI: 10.1175/AIES-D-23-0005.1 e230005

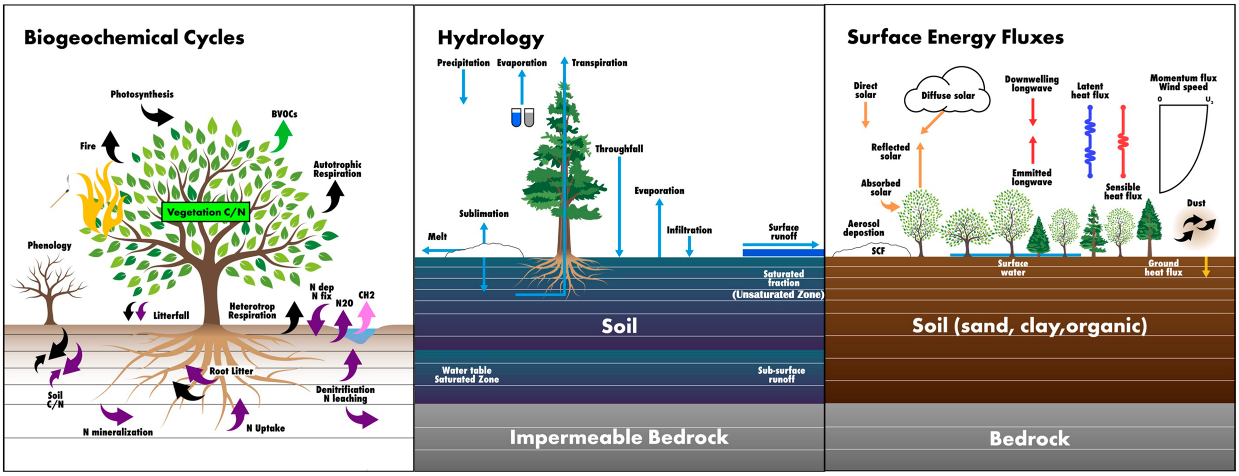

FIG. 1. Schematic representation of primary processes and functionality in land surface models, shown as an example of integrated ecohydrologic processes required for modeling across scales. For biogeochemical cycles, the black arrows denote carbon fluxes, and the purple arrows denote nitrogen fluxes. SCF is snow cover fraction, BVOC indicates biogenic volatile organic compounds, and C/N is the carbon-to-nitrogen ratio.

productivity (Scott et al. 2006), ecophysiology (Hultine et al. 2011; Vico et al. 2015), plant–soil interactions (Asbjornsen et al. 2011; Wang et al. 2019), and the biogeochemistry of terrestrial ecosystems (Kim et al. 2006; Wang et al. 2015) are all topics of interest in ecohydrology. These carbon, hydrologic, and energy cycle processes operate at scales ranging from stomates and microorganisms to canopies, watersheds, continents, and the entire globe (Fig. 1). Understanding interactions between important mechanisms across different scales is challenging. Constraining ecohydrologic models is limited by mechanistic knowledge gaps and by observations that are available at limited spatiotemporal scales of interest or at few sites (Rings et al. 2013; Massoud et al. 2019).

Artificial intelligence (AI) and machine learning (ML) approaches open up new possibilities for obtaining mechanistic insight from the diversity of data available at various scales. However, traditional hypothesis testing has been difficult to achieve with AI/ML models because much of the phenomenal success for some problems has been driven by “deep” neural network models that can be opaque and difficult to interpret in terms of process or underlying physical mechanism (Peters et al. 2007; Franz et al. 2010). Despite this obstacle, some inferences about the relative importance of different drivers with careful interpretation of variable importance estimates or model experiments have been made (Barnes et al. 2021). Generally, the terms “artificial intelligence” and “machine learning” are often used interchangeably; however, it is important to note that machine learning is a subset of the broader category of AI. For this paper, we refer to AI/ML methods as the broader category of using AI or ML strategies for ecohydrology.

Here we identify fundamental challenges in ecohydrology research across scales that can potentially be addressed by AI/ML approaches. We then provide a background of the state of the science in using AI/ML for ecohydrology. Following this, we highlight experimental, data and modeling opportunities that are available for this field. We then list the research priorities moving forward, and address the short-term, 5-yr, and 10-yr goals for using AI/ML in ecohydrology. Finally, we conclude with remarks on visionary perspectives for future ideas and potential research that can be achieved in ecohydrology.

2. Grand challenges

Various challenges are identified for using AI/ML in ecohydrology, which we have synthesized into three grand challenges: 1) accurate representation of complex ecohydrologic processes; 2) understanding and prediction of ecohydrologic responses to disturbance; and 3) machine learning applications for ecohydrology data and models.

The first of these grand challenges is to develop accurate and multiscale representations of land processes that incorporate heterogeneous patterns of water storage and fluxes, vegetation patterns and processes, water potential gradients, physiological function of vegetation, heterogeneous soil properties and processes, and biogeochemical cycling to understand and predict responses to climate change and climate extremes. Current models are starting to capture the necessary land processes to accurately simulate ecohydrologic processes at the plant scale (e.g., an increasing representation of plant hydrodynamics). For example, the Community Land Model (CLM; Lawrence et al. 2019) recently incorporated a submodule to simulate plant hydrodynamics mechanistically (Christoffersen et al. 2016; Kennedy et al. 2019), and the Bayesian-Based Carbon Data–Model Framework (CARDAMOM) now has a process-based hydrology model (Yang et al. 2022; Massoud et al. 2022) to simulate the hydrologic cycle and includes processes such as rooting depth and soil matric potential. Biological data for root network density and depth, root trait variability, and root responses to varying stresses are sparse and variable in space and time. Likewise, data on soil properties and processes are insufficient for constraining models. While aboveground processes are better understood, there are still significant gaps in the data needed to resolve species differences in ecophysiological processes, such as transfer of energy and mass between the atmosphere and the land surface, carbon assimilation and allocation, plant hydrodynamics, nutrient limitation, and the degree to which species-specific traits are plastic and adapted to local conditions (Fatichi et al. 2019; Xu and Trugman 2021). Moreover, traditional methods for integrating available data have been inadequate for developing insights into plant–soil interactions and how those interactions respond to and feed back on hydrology at the watershed, regional, and global scales (Ehrenfeld et al. 2005; Chen et al. 2021).

The second grand challenge identified is to develop models of pulse and press stresses, ecophysiological responses, and ecosystem structure and function to understand and predict ecosystem disturbances and recovery. Land surface models adequately simulate mean state behavior of vegetation, soils, and interactions with the atmosphere, but they often fail to capture responses to climate extremes either because of missing key processes or sensitivity to changes in temperature and precipitation that are too weak or too strong. Ecosystem disturbances and recovery patterns are especially challenging because traditional big-leaf models do not incorporate vegetation structural elements required to mechanistically account for changes in structure and function induced by climate, meteorological extremes, or biotic disturbance, like windthrow, fire, frost, drought, and insect or pathogen outbreaks (Seidl et al. 2011; Xu and Trugman 2021). Furthermore, biases in atmosphere forcing data, as well as scaling issues, can strongly affect land surface model ecohydrologic responses to climate extremes (Bonan et al. 2019). Models that do take into account disturbance processes and their relationships to dynamic ecosystem structure are still challenged in representing the complex processes that govern plant mortality or ecosystem assembly following disturbance, dramatically expanding the scope of model processes and thus increasing the complexity of the models (Liu et al. 2011; Huber et al. 2020).

The third and final grand challenge is to apply AI/ML to assimilate and calibrate emerging datasets, constrain model complexity, develop functional model benchmarks, and quantify the magnitude and sources of model and data uncertainty. A wide diversity of satellite and airborne remote sensing and in situ measurements are available to support ecohydrology research (e.g., Farella et al. 2022); however, these data are not well integrated and often do not include observations of variables needed by models (Karthikeyan et al. 2020). For example, Yan et al. (2023) demonstrated the hydrological sensitivity to parameter choices in a land surface model, and many of the parameters to which the model is sensitive are poorly constrained by observations because data are not well integrated (or spatially continuous) to support ecohydrology research. Furthermore, Yan et al. (2023) showed that hydrologic processes depend on a parameter defining the maximum storage of liquid water on leaf surface, but there is large uncertainty in this parameter, and the range of plausible values was determined by expert review of various individual peer reviewed papers. There are no integrated observational datasets that quantify this parameter to the extent required by land surface models.

To improve model performance, scientists tend to increase the complexity of models to capture processes for which there are insufficient measurements and highly uncertain parameters. AI/ML approaches are already being used to improve data through multisensor data fusion and quantitative methods for extrapolation and accounting for spatiotemporal heterogeneity (Shivaprakash et al. 2022). Similar approaches are showing promise for calibrating model parameters and quantifying model structural uncertainty (Dagon et al. 2020). AI/ML approaches are needed to further improve data, develop multivariate model benchmarks of functional performance, and constrain the ever-increasing complexity of models.

3. State of the science

In general, ecohydrology involves the coupling of soils, plants, and the atmosphere, requiring computationally intensive iterative solutions, which are difficult to integrate with limited observations. AI/ML approaches are already being used to 1) interpolate, extrapolate, and integrate data and models, accounting for nonlinear relationships among variables, to constrain and improve models (Mao et al. 2021); 2) build data-driven model components or parameterizations of processes from measurements and observational data products (Saunders et al. 2021); and 3) develop emulators and surrogate models of complex, nonlinear process representations for more efficient parameter estimation and optimization and model calibration (Massoud 2019; Dagon et al. 2020).

Bilinear interpolation, kriging, cluster analysis, random forests, model tree ensembles, convolutional neural networks, and other AI/ML methods have been applied to spatially sparse measurements to understand their representativeness (e.g., Hoffman et al. 2013; Kumar et al. 2016), to design optimal sampling networks (e.g., Keller et al. 2008; Hoffman et al. 2013; Vitharana et al. 2017), to analyze multidimensional model outputs (e.g., Braghiere et al. 2020; Burke et al. 2021), to constrain Earth system model projections (Yu et al. 2022), and to intelligently upscale and extrapolate environmental fluxes and characteristics over larger spatial domain (e.g., Langford et al. 2019; Steidinger et al. 2019; Jung et al. 2020; Konduri et al. 2020; Mishra et al. 2020; Barnes et al. 2021) using inferred relationships with environmental gradients, ecosystem dynamics, and remote sensing radiances.

Widening adoption of deep neural networks and the growth of meteorological and climate data have fueled interest in adopting AI/ML technologies for use in weather and climate models (Dueben and Bauer 2018). Leveraging the successes in rainfall prediction (Miao et al. 2015; Tao et al. 2016), soil moisture retrievals (Santi et al. 2016; Kolassa et al. 2017), and surface turbulent flux retrievals (Alemohammad et al. 2017; Jung et al. 2020; Braghiere et al. 2020), researchers are training deep neural networks as model parameterizations, initially for convection and subgrid-scale processes (Rasp et al. 2018; Gentine et al. 2018; Brenowitz and Bretherton 2018, 2019; Brenowitz et al. 2020), which are poorly captured by current models or are computationally prohibitive for decadal or longer-time-scale simulations.

Mimicking the response surface of model outputs using emulators or surrogate models has also become a useful AI/ML technique in ecohydrology. Massoud (2019) used polynomial chaos expansion emulators and sparse grid sampling for models of increasing complexity, including a 7-parameter hydrologic model, a 15-parameter ecohydrologic model, and an 81-parameter land surface model that was coupled to a vegetation dynamics model with ecosystem demography. Burke et al. (2021) employed random forests to identify the relative importance of biophysical and climatic parameters in predicting effects of fuel treatment in forests on forest dynamics. These researchers found that interactions between biophysical settings, climate, and fuel treatments are complex and have nonlinear effects on forest dynamics, water fluxes, and fire behavior. They further indicated that random forest models could be used to test additional scenarios without needing to run the complex model.

Overall, the types of AI/ML techniques for ecohydrologic applications are plentiful. Many works have emerged that make use of methods like deep neural networks, random forests, or hybrid AI/ML. For example, Saunders et al. (2021) uses many forms of random forest algorithms to improve on stomatal conductance estimates in comparison with older empirical approaches. In Aboelyazeed et al. (2023), a differentiable ecosystem modeling framework was introduced, which uses neural networks for photosynthesis simulations In ElGhawi et al. (2023), the authors combine physics-based modeling with AI/ML to infer stomatal resistances for hybrid modeling of evapotranspiration. So overall, many different applications of AI/ML in ecohydrology are possible. In the U.S. Department of Energy's (DOE's) Artificial Intelligence for Earth System Predictability (AI4ESP) workshop report (Hickmon et al. 2022), various AI/ML methods are mentioned and elaborated on. For instance, there is an entire chapter based on a session that investigated cross-cut AI/ML technologies.

4. New data and modeling opportunities

Advancing Earth system predictions with AI/ML methods requires large quantities of data regarding relevant processes across multiple spatial and temporal scales. Data requirements for training ML algorithms typically exceed the data needs for traditional process model development, verification, and validation. Therefore, additional data may be required from new laboratory and field measurements, manipulative experiments, airborne and satellite remote sensing, multisensor fusion and data synthesis, and modeling studies. Collecting, aggregating, distributing, and archiving these larger quantities of data and newly derived data products requires a systematic and organized approach to data management. Creating, finding, accessing, analyzing, visualizing, and utilizing these data to train ML algorithms necessitates an integrated storage and computational infrastructure available across projects, institutions, and individual investigators.

a. New data products

Ecohydrology research suffers from a lack of sufficient data across spatial scales from microbial and leaf scales to watershed and continental scales (e.g., Lin et al. 2023). In particular, because of its high spatial heterogeneity and difficulty in sampling, many more belowground data are needed to reduce characterization uncertainties and understand relationships between soil organic carbon and environmental factors (e.g., soil moisture, soil texture) that influence its formation and turnover and to understand root density, distribution and how roots change with environmental conditions. Similarly, species-specific plant data are needed globally to improve the representation of vegetation communities in models and better characterize and simulate responses to environmental change.

A wide variety of measurement techniques is required across scales, and a hierarchy of process-based and ML-based models is needed to simulate important processes across those scales and improve Earth system predictability (Fig. 2). For example, on the leaf scale, Yang et al. (2023) showed how the leaf angle is an important trait to consider in leaf-level photosynthesis representation. On the canopy scale, Lamour et al. (2023) investigated the effect of the vertical gradients within a canopy on the choice of photosynthetic parameters. On the global scale, Global Ecosystem Dynamics Investigation (GEDI) satellite data can be used for mapping forest canopy height globally (Potapov et al. 2021). AI/ML can be useful in acquiring such data through optimization of sampling or monitoring networks (e.g., Keller et al. 2008; Hargrove et al. 2003; Hoffman et al. 2013), autonomous control of measurement or sampling devices under changing conditions and extreme events, intelligent gap-filling and extrapolation of point measurements (e.g., Mishra et al. 2020; Jung et al. 2020), and fusion of data from different scales, multiple sensors, and in situ data from different agencies and measurement campaigns (e.g., Langford et al. 2017). New data products should be constructed in a manner that makes them easily accessible as a collection in standard, well-documented formats to facilitate ease of use and testing with a wide range of AI/ML approaches. One prominent example of such benchmark datasets, called ImageNet (Russakovsky et al. 2015), consists of a collection of images with associated labels (nouns) that can be used by the research community to train and test any number of object detection algorithms. ImageNet is, arguably, one of the major catalysts of the current AI/ML boom, and its impact has been such that various mainstream publications have written about it (e.g., https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world). Curated benchmark datasets like ImageNet for AI/ML in the Earth sciences can have a significant impact for ecohydrology and related topics.

Building collections of labeled Earth science data and offering them to the community would facilitate rapid testing of existing AI/ML methods and faster development of new AI/ML methods aimed specifically at addressing the needs of ecohydrology and related Earth science research. For example, Mishra et al. (2022) used a large number of field observations

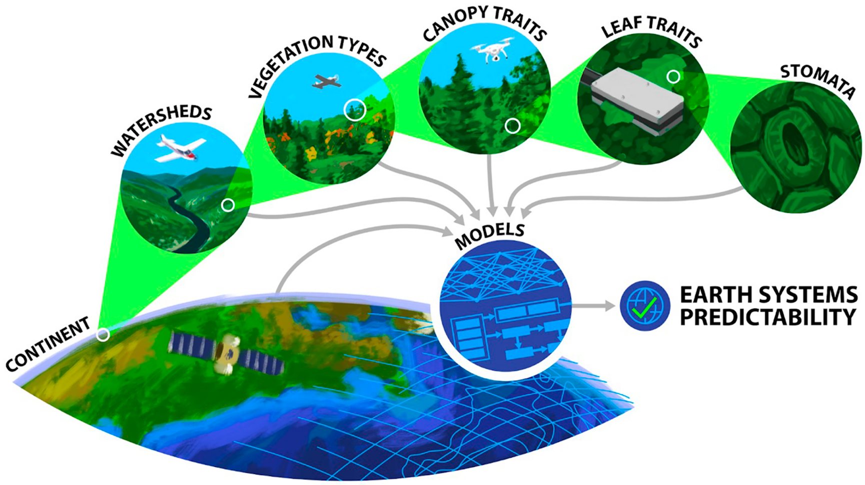

FIG. 2. A wide variety of measurement techniques are required across spatial scales from stomata to watersheds to improve representation of ecohydrologic processes with AI/ML approaches. These observations can be obtained by field and in situ measurements for small-scale processes, by aircraft sensors and drones for canopies and forests, and by remote sensing for larger watersheds and entire continents. This figure represents various aboveground processes and traits that can be measured, and there are also belowground processes not depicted here that can be main driving factors in ecohydrology.

and data of environmental factors, and derived the nonlinear relationships between environmental factors and soil organic carbon (SOC) stocks, which produced similar prediction accuracy as the AI/ML approaches. These mathematical relationships between environmental factors and SOC stocks can be used to benchmark environmental control representations of Earth system models.

Many data gaps and needs exist in the ecohydrology community, especially for AI/ML applications. According to the group of authors on this paper, it is not realistic to expect sufficient data to exist in the short to midterm to cover all data gaps needed for AI/ML in ecohydrology. With this paper, we are hoping to direct the community to identify the gaps where higher data needs exist, and to do so by using AI/ML technologies for ecohydrologic applications.

b. *Hybrid models*

Improving and developing new model parameterizations of ecohydrologic processes is inherent in the grand challenges presented above. However, where sufficient data are available, the opportunity exists to train deep neural networks for specific components within the model. Such efforts have begun, for hydrology (e.g., Slater et al. 2023) and for convection and subgrid-scale processes (Rasp et al. 2018; Gentine et al. 2018; Brenowitz and Bretherton 2018, 2019; Brenowitz et al. 2020), which are particularly suited to data-driven modeling approaches, as they are poorly captured by current process-based models or are computationally prohibitive for decadal or longer-time-scale simulations. Adding such capabilities in land surface models for simulating ecohydrologic processes could greatly advance the utility and performance of these models. Envisioned is a framework that employs such methods for data-driven, hybrid process-based/ML-based Earth system models (Schneider et al. 2017). As can be seen from this early work, lack of adequate data, numerical instabilities in coupling, and “out of sample” (i.e., “out of distribution” or extrapolative) problems must be overcome, but the outlook for these approaches is promising. Employing similar approaches for adding AI/ML capabilities in land surface models for simulating ecohydrologic processes could greatly advance the utility and performance of these models.

By explaining patterns identified by AI/ML, physically based models and observational datasets can be improved and optimized by incorporating missing processes, thereby providing transferability across space and time scales (Fig. 3). For example, AI/ML can be used to reduce the complexity of multidimensional outputs from physically based models (Massoud 2019), which would allow such models to be simulated with less complex input data. One could envision a framework that employs AI/ML methods for data-driven process representation alongside traditional differential equation-based representation of ecohydrologic processes, resulting in a hybrid process-based/ML-based model (Schneider et al. 2017; Braghiere et al. 2021; Wang et al. 2021a; Tsai et al. 2021; Feng et al. 2022, 2023). AI/ML can also be useful for the assimilation of data in process-based models, for example, by analyzing trait information carefully to determine prior distributions of parameters or for selecting sites that are representative of different data types for calibration. Furthermore, AI/ML can assist in the retrieval of remotely sensed information. To facilitate these visions, existing models must be made more modular so that individual process-based or ML-based parameterizations with a model may be swapped in and out as desired.

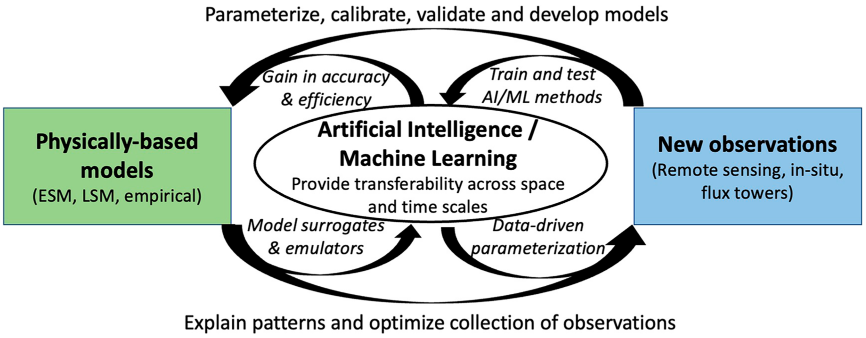
Parameterize, calibrate, validate and develop models
Gain in accuracy & efficiency
Train and test AI/ML methods
Physically-based models (ESM, LSM, empirical)
Artificial Intelligence / Machine Learning
Provide transferability across space and time scales
New observations (Remote sensing, in-situ, flux towers)
Model surrogates & emulators
Data-driven parameterization
Explain patterns and optimize collection of observations

FIG. 3. Combining physically based models and observed datasets with AI/ML methods enables identification of processes and patterns that can inform future model development and new observational campaigns. Such hybrid models provide transferability across space and time scales.

c. Computing and data infrastructure

The research community currently has access to high performance computing capabilities at large computing centers, such as the National Energy Research Scientific Computing Center (NERSC) at Lawrence Berkeley National Laboratory, the Oak Ridge Leadership Computing Facility (OLCF), and the Argonne Leadership Computing Facility within the DOE. The community has access to large collections of data at data centers, like DOE's Atmospheric Radiation Measurement (ARM) Data Center (ADC), Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE), Earth System Grid Federation (ESGF), NASA's Distributed Active Archive Centers (DAACs), Consortium of Universities for the Advancement of Hydrologic Science's (CUAHSI) Hydroshare, and others. These data centers operate as stand-alone resources and require data users to download data to their own computational resources. This process of downloading data, preprocessing and integrating the data, and then performing simulations and analysis is tedious and unnecessary given recent technological developments. When developing and deploying AI/ML methods, the difficulty of this workflow will increase since high-speed access to vastly larger data collections will be required for training AI/ML models, potentially doing such training as part of simulation itself.

There are efforts in the Earth science community to develop and implement cloud-based virtual centers for computing and data needs. For example, the DOE is seeking input on the need and the structure of a unified data framework that links or integrates existing data activities for next-generation data management and analysis. However, there have not been any concrete conversations or plans regarding a cloud-based virtual center. This could become possible by vendors such as Amazon Web Services, yet these types of efforts have not been initiated yet. Importantly, there is a distinction between commercial clouds and the combination of cloud-based deployments and cloud-based technologies used for research (e.g., Jupyter hubs and S3 buckets), which do not necessarily have to be on commercial clouds. For example, open research clouds exist, which could be a valuable platform to conduct the types of efforts outlined in this paper.

5. Research priorities

Priorities for near-term research in ecohydrology should aim to prepare the research community to address the grand challenges listed above. This includes improving characterization of soil and vegetation properties, improving representation of water stores and fluxes, developing models of extremes and ecosystem disturbance and recovery, and developing new assimilation and analysis capabilities to help constrain models and quantify sources of both model and data uncertainty. The research community is at a stage where progress can be made in creating benchmark "AI ready" datasets and developing initial AI/ML parameterizations and process emulators. Initial research and development activities should engage a broader, more multidisciplinary, community of researchers, particularly in mathematics and computer science. Transitioning the community to significant use of AI/ML approaches in ecohydrology and climate science will require enhanced efforts to train the next generation of researchers to use new tools and methods. National scientific workforce development activities should consider how best to deliver the additional knowledge and training to early career scientists.

a. Benchmark datasets

While Earth system and environmental data centers distribute and archive a wide variety of data collections from in situ measurements, monitoring networks, and airborne and satellite remote sensing platforms, they do not typically lead activities to synthesize data products across those collections for specific research purposes. Instead, funded or volunteer working groups are often formed to synthesize data to address specific science questions or hypotheses. Such working groups may be catalyzed by existing projects [e.g., Reducing Uncertainties in Biogeochemical Interactions through Synthesis and Computation (RUBISCO) working groups on Soil Carbon Dynamics, AmeriFlux, and Soil Moisture], data collection activities or databases [e.g., various AmeriFlux and Fluxnet working groups, International Soil Carbon Network (ISCN), International Soil Radiocarbon Database (ISRaD), "TRY" Plant Trait Database, Fine-Root Ecology Database (FRED), National Ecological Observatory Network (NEON), International

Soil Moisture Network], or data synthesis centers [e.g., National Center for Ecological Analysis and Synthesis (NCEAS), National Institute for Mathematical and Biological Synthesis (NIMBioS), Powell Center for Analysis and Synthesis, Critical Zone Collaborative Network (CZCN),CUAHSI, Aspen Global Change Institute] sponsored by the National Science Foundation, U.S. Geological Survey, U.S. Department of Agriculture, and other agencies and nongovernmental organizations. However, because these working group activities are often narrowly focused, they may not produce synthesized data products that are of general use, well-documented, easily distributed, archived, and maintained over time. A more systematic approach with a broader vision for reusability and maintainability is required to generate benchmark datasets for training, testing, and benchmarking AI/ML models.

Producing and maintaining large collections of understandable and reusable data, like that from ImageNet (Russakovsky et al. 2015), will be of great utility to the ecohydrology research community and will facilitate wider engagement of the mathematics and computer science communities already involved in developing and applying AI/ML methods. Some of these datasets will be similar to climate reanalysis data products (e.g., ERA5; Hersbach et al. 2020), synthesized data either used for model evaluation by software like the International Land Model Benchmarking (ILAMB) package (Collier et al. 2018) or global/large-scale process investigation (Sprenger et al. 2021), or satellite-based remote sensing data products (Dalla Mura et al. 2015; Hong et al. 2021; Potapov et al. 2021). Such datasets must be highly multivariate for AI/ML methods to uncover relationships, integrated in a consistent manner for direct use without translation or conversion, available across multiple spatial and temporal scales, and contain long time series of a large number of samples, points, or grid cells. Furthermore, such datasets should draw upon many independent data sources, such as data fused from multiple remote sensing platforms, and be calibrated with in situ measurements and continental-scale monitoring networks (e.g., Wang et al. 2021b). To be of greatest utility, these data must be maintained and distributed by existing or new data centers, and integrated computing and storage infrastructure should be developed to facilitate data discovery and eliminate barriers to data movement and download.

b. Hybrid modeling

Given the availability of growing volumes of observational data and in situ measurements, the Earth system modeling community is beginning to adopt data-driven approaches for high resolution weather and climate simulations (Schneider et al. 2017). A ML framework could be used to integrate the wealth of leaf-level fluorescence and gas exchange measurements (e.g., Leafweb), AmeriFlux and Flux Network (FLUXNET) ecosystem fluxes, and Free Air Carbon Dioxide Enrichment (FACE) and Spruce and Peatland Responses Under Changing Environments (SPRUCE) data to develop a unified treatment of stomatal responses, assimilation, and acclimation to changes in environmental conditions like hydrology or soil moisture. AI/ML-based models of stomatal conductance and plant hydrodynamics should be employed to produce a hybrid process-based/ML-based land model with the aim of reducing the uncertainty in soil moisture and carbon assimilation. Such models can incorporate as many processes as possible and as a result can have extremely high dimensionality (Fig. 4) or alternatively can have more simple versions with lower dimensionality and complexity depending on what is needed for the specific application (Fig. 5). Such hybrid ecohydrology models could also inform watershed models to deliver dynamic ecological process representations, such as the EcH2O model (Maneta and Silverman 2013) that captures the biophysical dynamics of vegetation and the hydrologic cycle at the watershed scale, but such information is often absent in these models. In addition, ML models can be developed to improve the characterization of soil organic carbon and soil bulk properties to further reduce soil moisture uncertainties. ML methods should be explored to scale leaf-level and ecosystem processes to the watershed scale for seasonal-to-interannual predictions, through a hierarchy of ML and process-based models, and further to regional and continental scales for interannual-to-decadal predictions. For research questions involving disturbance and recovery, new mechanistic modeling approaches (e.g., Hanan et al. 2021) are advancing our understanding, and these models would benefit from detailed information about changing vegetation structure to support both model parameterization and evaluation. Modeling disturbance is an area, due to its complexity, that would particularly benefit from hybrid approaches since AI/ML methods can fill in some of the knowledge gaps in simulating these processes.

Under expected future climate change, plant and soil processes could experience new conditions that have not yet been seen by existing observational data. For example, it is not yet known how the impacts of climate change will affect things such as the distributions of vegetation as a result of novel or disappearing climates (Williams et al. 2007), how changing climate will impact ecosystems at different spatial scales (Maclean 2020), or how global climate change such as intensifying droughts will affect forest ecosystems and their microbiomes across different climatic zones (Baldrian et al. 2023). Thus, experimental data based on analogs to potential future climate conditions such as FACE and SPRUCE data should be used to develop climate-adaptive ML models for the processes described above. This approach could enable significant steps forward in developing and integrating new and alternative parameterizations within Earth system models to produce a hybrid process-based/ML modeling framework (Reichstein et al. 2019). The requirement for reducing uncertainties in ecohydrologic processes dictates prioritizing process representations of land–atmosphere interactions (energy, water, carbon, and nutrients) that are 1) highly uncertain but for which observational data are available and 2) computationally expensive. Measurements of leaf-level responses to environmental variations can be related to measurements made at the canopy scale to reduce uncertainties in canopy integration schemes. AI/ML methods can be applied to scale up plant responses—informed by ecosystem- and watershed-scale measurements, upscaled soil properties, and remote

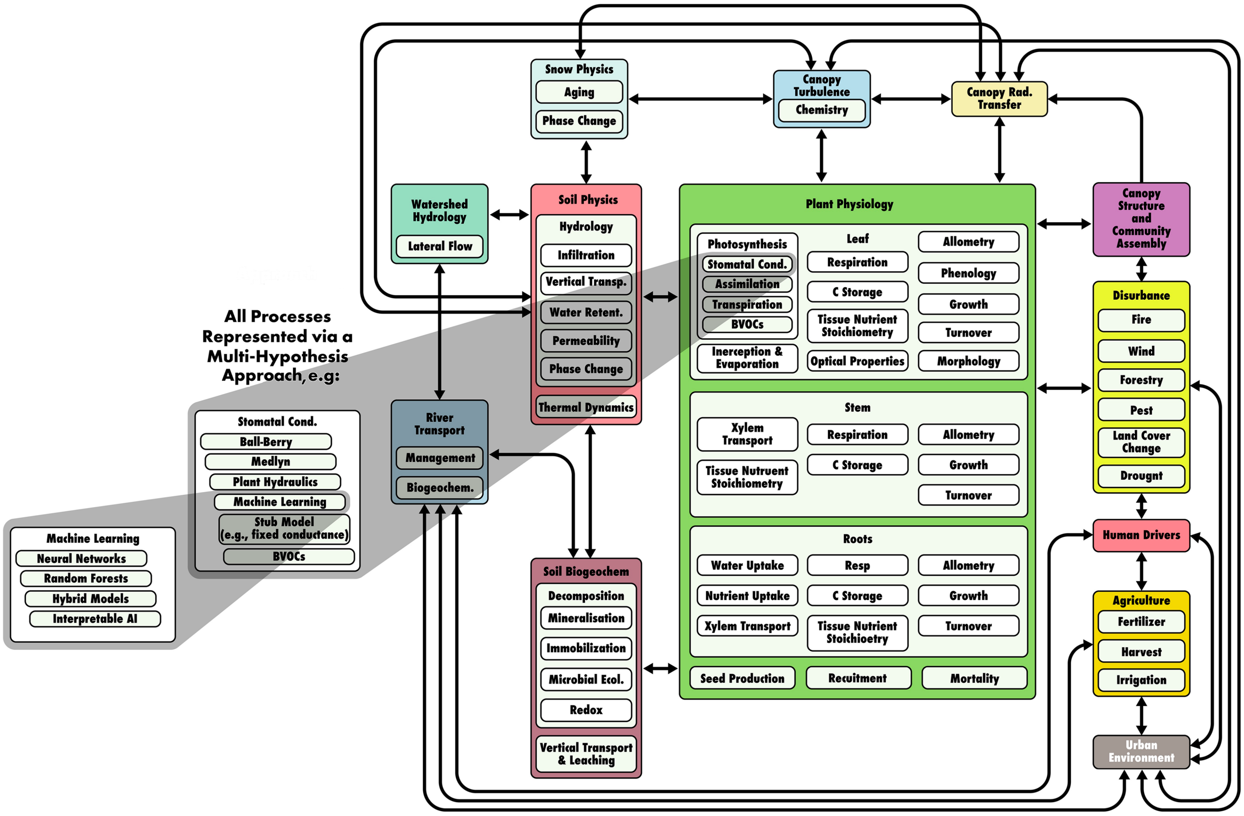

FIG. 4. A process schematic of a full-complexity land surface model. Processes, and sets of processes, are represented as boxes in the diagram, with information connections represented as arrows. All processes—although here shown only for stomatal conductance—are intended to allow alternative specification, including possibly multiple hypothetical process realizations, empirical or machine learning–derived formulations, and/or simplified stub or null representations to allow for holding a given process constant while other processes vary (adapted from Fisher and Koven 2020).

sensing data—to bound water budgets for watersheds (e.g., Massoud et al. 2022) and quantify risks of flooding and drought, particularly under water cycle extremes. While the primary motivation is to improve mechanistic understanding of these processes across scales, by connecting a chain of hierarchical AI/ML-empowered models to weather forecasting systems, the results may be useful for informing probabilistic risk analysis to quantify risks for urban areas and other built infrastructure and to better quantify drought impacts on streamflow for energy and water utilities.

One of the challenges of hybrid models and increasingly complex ecohydrologic models in general is the opacity of these modeling frameworks. For users to appropriately apply these models and gain mechanistic understanding requires that the underlying assumptions and process representations be visible (Tague and Frew 2021). Advances in model documentation and visualization of outputs and of underlying model architecture and performance will be needed to address this challenge (e.g., National Water Model; Wan et al. 2022).

c. *Multidisciplinary engagement with AI/ML researchers*

New research in ecohydrology employing AI/ML approaches will benefit from strong collaboration with scientists in mathematics and computer science, who routinely apply such methods in other disciplines and who are actively developing new methods specific to research needs in other domains (Rolnick et al. 2022; Hickmon et al. 2022; Sukanya and Joseph 2023). For example, the review shown in Rolnick et al. (2022) encompasses exciting research questions as well as promising business opportunities for the AI/ML community, and the authors call on the AI/ML community to join the global effort against climate change. The paper by Rolnick et al. (2022) is primarily written by ML experts, and documents how their community can contribute to climate mitigation and adaptation, and Earth system prediction efforts. Strengthening such collaborations will require frequent interaction between domain experts and computer scientists, mathematical generalization of specific process representations in models, and well-documented benchmark datasets. For ecohydrology, engaging with mathematicians and computer scientists will enable leveraging of research and development activities already underway, and it will foster long-lasting collaborations that will benefit both sets of communities. For long lasting changes fostering intense cross-disciplinary collaborations, mathematics and computer science should become more prominent in Earth system science education at both the undergraduate and graduate levels.

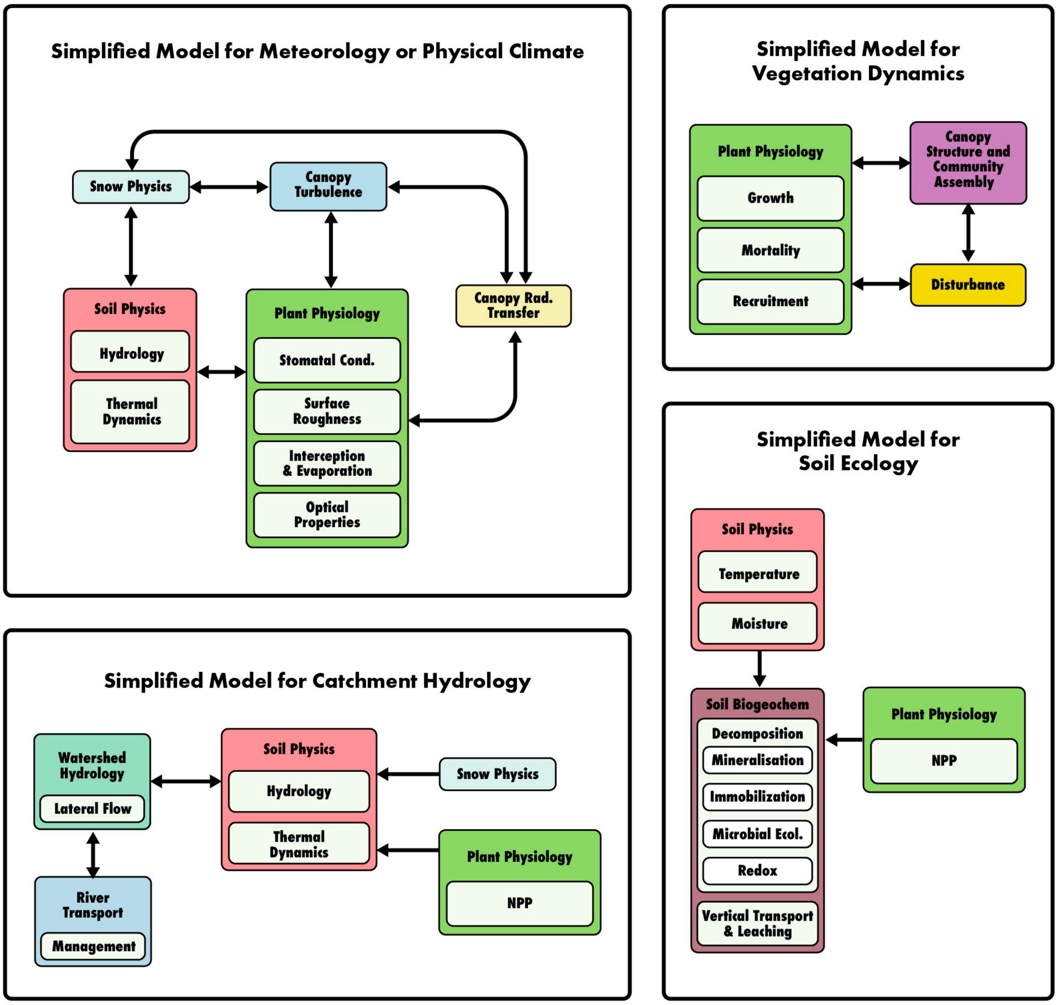

FIG. 5. As in Fig. 4, but for some possible simplified configurations of a land surface model. Simplifications shown here are formatted for different purposes, such as meteorology or physical climate, vegetation dynamics, catchment hydrology, or soil ecology (adapted from Fisher and Koven 2020).

d. Integrated data and computational research infrastructure

Adding AI/ML approaches for data acquisition, processing, assimilation, modeling, and analysis will require improved infrastructure for large datasets and high performance computational capacity targeted for AI/ML. The growth opportunity is to build integrated computing and data infrastructure that eliminates the challenges of finding, acquiring, and downloading data. Benchmark AI/ML data should be accessible from all large computing environments, no matter where those data reside or are archived. This could be accomplished through application programming interfaces and data transport services, like Globus (https://globus.org/), that hide the details of data movement and exploit high bandwidth networks to deliver data as needed for simulation and analysis. Funding agencies might coordinate in the creation of a model-data integration center that could provide such integrated storage and computing resources for the growing Earth system science community. The center could provide data hosting services, offer compute-near-the-data computational infrastructure and "AI/ML as a service" capabilities, and sponsor training activities and multidisciplinary working groups focused on improving new or advanced research topics that may have some element of risk. Such a center could lower the bar of entry for laboratory and university scientists, fostering multidisciplinary engagement, while enabling research with tools not otherwise easily accessible or usable.

e. Training and workforce development

To advance research with AI/ML approaches, current and next generation researchers need training on the wide variety of AI/ML methods, data management, large-scale analytics techniques, and use of integrated computational and data resources. This could be accomplished through fellowships that support national laboratory internships for promising graduate students, training courses for postdoctoral and early career scientists (akin to open access online classes for hydrology at https://www.hydrolearn.org/), and seminars and hackathons for existing staff (like hydrology seminars provided by the CUAHSI Community at https://www.cuahsi.org/community). These activities could begin with webinars that highlight existing research in national laboratories and universities and virtual hackathons that demonstrate analysis techniques, useful software packages, and strategies for applying emerging datasets. These education and training activities should be an integrated part of training of the next generation workforce with diverse research scientists to meet the needs of the nation.

6. Short-, mid-, and long-term goals

In this section, we identify specific goals for each research priority. Incremental progress through the following goals is expected to reduce model uncertainties and improve prediction accuracies leading to actionable science outcomes.

a. *Short-term (<5 years) goals*

- Develop a collection of "AI ready" benchmark datasets for leaf-level measurements of fluxes of energy, water, carbon, and nutrients; canopy-level observations of evapotranspiration and productivity; and continental-scale estimates of carbon and water cycle time series from in situ measurements and airborne and satellite remote sensing.
- Synthesize existing data in a network-of-networks approach to provide AI-ready datasets on subsurface characterization (e.g., high-frequency soil moisture dynamics, soil water tracer data) across large environmental gradients to study the soil-plant feedbacks.
- Improve the modularity of current models so that individual process-based parameterizations can be isolated and swapped with AI/ML-based versions of parameterizations.
- Develop an initial set of AI/ML-based parameterizations for photosynthesis, stomatal conductance, and other vegetation and soil processes that can be integrated as components into hybrid models.
- Establish collaborative opportunities across Earth system science, mathematics, statistics, and computer science directed at developing and applying novel and domain-specific AI/ML methods to improve accuracy of ecohydrology process representations in Earth system models.
- Design and begin implementation of an integrated data and computational infrastructure to support AI/ML in Earth system science. This could leverage existing data centers, computational centers, and software infrastructure, and potentially be transitioned to its own center or facility for broader engagement of the research community.
- Initiate a webinar series for educating and training cross-disciplinary researchers across career stages about the use of AI/ML methods and tools. Conduct virtual and in-person hackathons for more rigorous training of graduate students, postdoctoral scholars, and early career scientists.

b. *Mid-term (5 years) goals*

- Develop an initial modeling framework for swapping or interchanging process-based and AI/ML-based parameterizations within Earth system models.
- Foster cross-disciplinary research and training by sponsoring transdisciplinary working groups that include observational scientists, modelers, data scientists, mathematicians, and computer scientists to take advantage of the benchmark data, AI/ML model frameworks, and integrated computational and storage resources to address specific science questions in ecohydrology.
- Develop accurate and efficient science-guided AI/ML systems or models to predict effects of different ecohydrologic disturbances and postdisturbance responses and feedbacks.
- Employ AI/ML to generate new synthetic data for training AI/ML algorithms, for example, photographing each root core collected and developing an AI/ML algorithm to help understand and fast track improvements in data observations of this kind.
- Develop AI/ML algorithms that can infer what additional measurements are needed and what optimal sampling frequencies and spatial distributions will lead to improvements in ecohydrology models.

c. *Long-term (10 years) goals*

- Deploy a fully functioning modeling framework for easily configuring and monitoring AI/ML-based parameterizations alongside process-based parameterizations within Earth system models, supporting online training and in situ analysis and visualization.
- Deploy a fully functioning explainable AI/ML framework that can identify where to collect data (space/time gaps), what processes need to be improved (physics/chemistry/biology gaps), and how to better manage and analyze data for ecohydrologic applications.
- Deploy a fully functioning AI/ML-based ecohydrologic subsystem for Earth system models that is tested and calibrated for accurate predictions across relevant space and time scales and that includes ecosystem disturbance and recovery process representations.
- Establish a multiagency AI/ML center to provide computational and storage infrastructure, necessary benchmark data, a wide variety of models at different scales, software tools for analysis and visualization, and staff to support a collection of working groups that have proposed to address key science questions in ecohydrologic predictability.

7. Visionary perspectives for future ideas and potential research

The evolution of AI/ML informed Earth system and multiscale ecohydrologic models will require a community effort, involving multiple disciplines, advanced training, and new ways of designing, implementing, parameterizing, and communicating model output for understanding and for solving environmental problems. The collective teamwork required may ultimately need new ways of working together, creating new incentive structures to promote collaboration and communication. Novel ways are needed to learn not only how to model, but how to effectively collaborate on AI/ML applications. New university doctoral programs can be envisioned that are more closely aligned with multidisciplinary research laboratories that focus on collaborative AI/ML research. We expect that expanding multidisciplinary training and cultivating multidisciplinary collaboration between ecohydrologists and AI/ML experts will allow AI/ML to drive advancements in ecohydrology even beyond those envisioned here by our research community.

New efforts aimed at building flexible model structures, such as the Climate Modeling Alliance (CliMA) for example, can employ AI/ML methods for data-driven process representation and can leverage recent advances in the computational and data sciences, to learn directly from a wealth of Earth observations from space and the ground. These types of efforts can develop an Earth system or ecohydrologic model that automatically learns from diverse data sources and exploits advances in AI/ML to learn from observations and from data generated on demand in targeted high-resolution simulations (Schneider et al. 2017; Braghiere et al. 2021; Wang et al. 2021a).

A new type of modeling framework called digital twins is also a topic of current and future research for the ecohydrologic sciences. In essence, a digital twin is a virtual model of a physical object, and spans the object's life cycle and uses real-time data sent from sensors on the object to simulate its behavior. For example, the European Union plans to fund the development of digital twins of Earth, with aims for these "twins" to be more than big data atlases, and rather creating a qualitatively new Earth system simulation and observation capability (Bauer et al. 2021). It is envisioned that a digital twin of Earth can act as an information system that exposes users to a digital replication of the state and temporal evolution of the Earth system constrained by available observations and the laws of physics. For practical reasons, the "digital twins of the Earth" may generate the actionable intelligence that is necessary to address global change challenges (Nativi et al. 2021). Setting up digital ecosystems to create digital twins for ecohydrologic applications may be an additional path forward for fusing AI/ML algorithms, process-based models, and the wealth of observations available to gain understanding and make predictions of ecohydrologic systems.

Acknowledgments. The authors acknowledge all of the efforts made as part of the Artificial Intelligence for Earth System Predictability (AI4ESP) workshop. This research was partially supported by the RUBISCO Science Focus Area (RUBISCO SFA KP1703), which is sponsored by the Regional and Global Model Analysis (RGMA) activity of the Earth and Environmental Systems Modeling (EESM) Program in the Earth and Environmental Systems Sciences Division (EESSD) of the Office of Biological and Environmental Research (BER) in the U.S. Department of Energy Office of Science. This paper has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy (DOE). The U.S. government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this paper, or allow others to do so, for U.S. government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan. Part of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Government sponsorship is acknowledged. The authors declare no conflicts of interest. All authors contributed to the discussions that inspired this paper and were responsible for writing parts of the paper in the early stages of preparation. Author Massoud wrote the first complete draft, and all authors contributed to writing and editing the current version of the paper.

Data availability statement. There were no specific datasets used in this study.

REFERENCES

Aboelyazeed, D., C. Xu, F. M. Hoffman, J. Liu, A. W. Jones, C. Rackauckas, K. Lawson, and C. Shen, 2023: A differentiable, physics-informed ecosystem modeling and learning framework for large-scale inverse problems: Demonstration with photosynthesis simulations. *Biogeosciences*, **20**, 2671–2692, https://doi.org/10.5194/bg-20-2671-2023.

Alemohammad, S. H., and Coauthors, 2017: Water, Energy, and Carbon with Artificial Neural Networks (WECANN): A statistically based estimate of global surface turbulent fluxes and gross primary productivity using solar-induced fluorescence. *Biogeosciences*, **14**, 4101–4124, https://doi.org/10.5194/bg-14-4101-2017.

Asbjornsen, H., and Coauthors, 2011: Ecohydrological advances and applications in plant–water relations research: A review. *J. Plant Ecol.*, **4**, 3–22, https://doi.org/10.1093/jpe/rtr005.

Baldrian, P., R. López-Mondéjar, and P. Kohout, 2023: Forest microbiome and global change. *Nat. Rev. Microbiol.*, **21**, 487–501, https://doi.org/10.1038/s41579-023-00876-4.

Barnes, M. L., and Coauthors, 2021: Improved dryland carbon flux predictions with explicit consideration of water-carbon coupling. *Commun. Earth Environ.*, **2**, 248, https://doi.org/10.1038/s43247-021-00308-2.

Bauer, P., B. Stevens, and W. Hazeleger, 2021: A digital twin of Earth for the green transition. *Nat. Climate Change*, **11**, 80–83, https://doi.org/10.1038/s41558-021-00986-y.

Bonan, G. B., D. L. Lombardozzi, W. R. Wieder, K. W. Oleson, D. M. Lawrence, F. M. Hoffman, and N. Collier, 2019: Model structure and climate data uncertainty in historical simulations of the terrestrial carbon cycle (1850–2014). *Global Biogeochem. Cycles*, **33**, 1310–1326, https://doi.org/10.1029/2019GB006175.

Braghiere, R. K., M. A. Yamasoe, N. M. Évora do Rosário, H. Ribeiro da Rocha, J. de Souza Nogueira, and A. C. de Araújo, 2020: Characterization of the radiative impact of aerosols on CO_2 and energy fluxes in the Amazon deforestation arch using artificial neural networks. *Atmos. Chem. Phys.*, **20**, 3439–3458, https://doi.org/10.5194/acp-20-3439-2020.

——, and Coauthors, 2021: Accounting for canopy structure improves hyperspectral radiative transfer and sun-induced chlorophyll fluorescence representations in a new generation Earth system model. *Remote Sens. Environ.*, **261**, 112497, https://doi.org/10.1016/j.rse.2021.112497.

Brenowitz, N. D., and C. S. Bretherton, 2018: Prognostic validation of a neural network unified physics parameterization. *Geophys. Res. Lett.*, **45**, 6289–6298, https://doi.org/10.1029/2018GL078510.

——, and ——, 2019: Spatially extended tests of a neural network parameterization trained by coarse-graining. *J. Adv. Model. Earth Syst.*, **11**, 2728–2744, https://doi.org/10.1029/2019MS001711.

——, T. Beucler, M. Pritchard, and C. S. Bretherton, 2020: Interpreting and stabilizing machine-learning parameterizations of convection. *J. Atmos. Sci.*, **77**, 4357–4375, https://doi.org/10.1175/JAS-D-20-0082.1.

Burke, W. D., C. Tague, M. C. Kennedy, and M. A. Moritz, 2021: Understanding how fuel treatments interact with climate and biophysical setting to affect fire, water, and forest health: A process-based modeling approach. *Front. For. Global Change*, **3**, 591162, https://doi.org/10.3389/ffgc.2020.591162.

Chen, X., and Coauthors, 2021: Integrating field observations and process-based modeling to predict watershed water quality under environmental perturbations. *J. Hydrol.*, **602**, 125762, https://doi.org/10.1016/j.jhydrol.2020.125762.

Christoffersen, B. O., and Coauthors, 2016: Linking hydraulic traits to tropical forest function in a size-structured and trait-driven model (TFS v.1-hydro). *Geosci. Model Dev.*, **9**, 4227–4255, https://doi.org/10.5194/gmd-9-4227-2016.

Collier, N., F. M. Hoffman, D. M. Lawrence, G. Keppel-Aleks, C. D. Koven, W. J. Riley, M. Mu, and J. T. Randerson, 2018: The International Land Model Benchmarking (ILAMB) system: Design, theory, and implementation. *J. Adv. Model. Earth Syst.*, **10**, 2731–2754, https://doi.org/10.1029/2018MS001354.

Dagon, K., B. M. Sanderson, R. A. Fisher, and D. M. Lawrence, 2020: A machine learning approach to emulation and biophysical parameter estimation with the Community Land Model, version 5. *Adv. Stat. Climatol. Meteor. Oceanogr.*, **6**, 223–244, https://doi.org/10.5194/ascmo-6-223-2020.

Dalla Mura, M., S. Prasad, F. Pacifici, P. Gamba, J. Chanussot, and J. A. Benediktsson, 2015: Challenges and opportunities of multimodality and data fusion in remote sensing. *Proc. IEEE*, **103**, 1585–1601, https://doi.org/10.1109/JPROC.2015.2462751.

Dueben, P. D., and P. Bauer, 2018: Challenges and design choices for global weather and climate models based on machine learning. *Geosci. Model Dev.*, **11**, 3999–4009, https://doi.org/10.5194/gmd-11-3999-2018.

Ehrenfeld, J. G., B. Ravit, and K. Elgersma, 2005: Feedback in the plant-soil system. *Annu. Rev. Environ. Resour.*, **30**, 75–115, https://doi.org/10.1146/annurev.energy.30.050504.144212.

ElGhawi, R., B. Kraft, C. Reimers, M. Reichstein, M. Körner, P. Gentine, and A. J. Winkler, 2023: Hybrid modeling of evapotranspiration: Inferring stomatal and aerodynamic resistances using combined physics-based and machine learning. *Environ. Res. Lett.*, **18**, 034039, https://doi.org/10.1088/1748-9326/acbbe0.

Farella, M. M., J. B. Fisher, W. Jiao, K. B. Key, and M. L. Barnes, 2022: Thermal remote sensing for plant ecology from leaf to globe. *J. Ecol.*, **110**, 1996–2014, https://doi.org/10.1111/1365-2745.13957.

Fatichi, S., C. Pappas, J. Zscheischler, and S. Leuzinger, 2019: Modelling carbon sources and sinks in terrestrial vegetation. *New Phytol.*, **221**, 652–668, https://doi.org/10.1111/nph.15451.

Feng, D., J. Liu, K. Lawson, and C. Shen, 2022: Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy. *Water Resour. Res.*, **58**, e2022WR032404, https://doi.org/10.1029/2022WR032404.

——, H. Beck, K. Lawson, and C. Shen, 2023: The suitability of differentiable, learnable hydrologic models for ungauged regions and climate change impact assessment. *Hydrol. Earth Syst. Sci.*, **27**, 2357–2373, https://doi.org/10.5194/hess-27-2357-2023.

Fisher, R. A., and C. D. Koven, 2020: Perspectives on the future of land surface models and the challenges of representing complex terrestrial systems. *J. Adv. Model. Earth Syst.*, **12**, e2018MS001453, https://doi.org/10.1029/2018MS001453.

Franz, T. E., K. K. Caylor, J. M. Nordbotten, I. Rodríguez-Iturbe, and M. A. Celia, 2010: An ecohydrological approach to predicting regional woody species distribution patterns in dryland ecosystems. *Adv. Water Resour.*, **33**, 215–230, https://doi.org/10.1016/j.advwatres.2009.12.003.

Gentine, P., M. Pritchard, S. Rasp, G. Reinaudi, and G. Yacalis, 2018: Could machine learning break the convection parameterization deadlock? *Geophys. Res. Lett.*, **45**, 5742–5751, https://doi.org/10.1029/2018GL078202.

Guswa, A. J., and Coauthors, 2020: Advancing ecohydrology in the 21st century: A convergence of opportunities. *Ecohydrology*, **13**, e2208, https://doi.org/10.1002/eco.2208.

Hanan, E. J., and Coauthors, 2021: How climate change and fire exclusion drive wildfire regimes at actionable scales. *Environ. Res. Lett.*, **16**, 024051, https://doi.org/10.1088/1748-9326/abd78e.

Hargrove, W. W., F. M. Hoffman, and B. E. Law, 2003: New analysis reveals representativeness of the AmeriFlux network. *Eos, Trans. Amer. Geophys. Union*, **84**, 529–535, https://doi.org/10.1029/2003EO480001.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Hickmon, N. L., C. Varadharajan, F. M. Hoffman, S. Collis, and H. M. Wainwright, 2022: Artificial Intelligence for Earth System Predictability (AI4ESP): 2021 Workshop Report. Tech. Rep. ANL-22/54 177828, 413 pp., https://doi.org/10.2172/1888810.

Hoffman, F. M., J. Kumar, R. T. Mills, and W. W. Hargrove, 2013: Representativeness-based sampling network design for the State of Alaska. *Landscape Ecol.*, **28**, 1567–1586, https://doi.org/10.1007/s10980-013-9902-0.

Hong, D., J. Hu, J. Yao, J. Chanussot, and X. X. Zhu, 2021: Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model. *ISPRS J. Photogramm. Remote Sens.*, **178**, 68–80, https://doi.org/10.1016/j.isprsjprs.2021.05.011.

Huber, N., H. Bugmann, and V. Lafond, 2020: Capturing ecological processes in dynamic forest models: Why there is no silver bullet to cope with complexity. *Ecosphere*, **11**, e03109, https://doi.org/10.1002/ecs2.3109.

Hultine, K. R., and S. E. Bush, 2011: Ecohydrological consequences of non-native riparian vegetation in the southwestern United States: A review from an ecophysiological perspective. *Water Resour. Res.*, **47**, W07542, https://doi.org/10.1029/2010WR010317.

Jung, M., and Coauthors, 2020: Scaling carbon fluxes from eddy covariance sites to globe: Synthesis and evaluation of the FLUXCOM approach. *Biogeosciences*, **17**, 1343–1365, https://doi.org/10.5194/bg-17-1343-2020.

Karthikeyan, L., I. Chawla, and A. K. Mishra, 2020: A review of remote sensing applications in agriculture for food security: Crop growth and yield, irrigation, and crop losses. *J. Hydrol.*, **586**, 124905, https://doi.org/10.1016/j.jhydrol.2020.124905.

Keller, M., D. S. Schimel, W. W. Hargrove, and F. M. Hoffman, 2008: A continental strategy for the National Ecological Observatory Network. *Front. Ecol. Environ.*, **6**, 282–284, https://doi.org/10.1890/1540-9295(2008)6[282:ACSFTN]2.0.CO;2.

Kennedy, D., S. Swenson, K. W. Oleson, D. M. Lawrence, R. Fisher, A. C. Lola da Costa, and P. Gentine, 2019: Implementing plant hydraulics in the Community Land Model, version 5. *J. Adv. Model. Earth Syst.*, **11**, 485–513, https://doi.org/10.1029/2018MS001500.

Kim, J., and Coauthors, 2006: HydroKorea and CarboKorea: Cross-scale studies of ecohydrology and biogeochemistry in a heterogeneous and complex forest catchment of Korea. *Ecol. Res.*, **21**, 881–889, https://doi.org/10.1007/s11284-006-0055-3.

Kolassa, J., R. H. Reichle, and C. S. Draper, 2017: Merging active and passive microwave observations in soil moisture data assimilation. *Remote Sens. Environ.*, **191**, 117–130, https://doi.org/10.1016/j.rse.2017.01.015.

Konduri, V. S., J. Kumar, W. W. Hargrove, F. M. Hoffman, and A. R. Ganguly, 2020: Mapping crops within the growing season across the United States. *Remote Sens. Environ.*, **251**, 112048, https://doi.org/10.1016/j.rse.2020.112048.

Kumar, J., F. M. Hoffman, W. W. Hargrove, and N. Collier, 2016: Understanding the representativeness of FLUXNET for upscaling carbon flux from eddy covariance measurements. *Earth Syst. Sci. Data Discuss.*, https://doi.org/10.5194/essd-2016-36, withdrawn preprint.

Lamour, J., and Coauthors, 2023: The effect of the vertical gradients of photosynthetic parameters on the CO_2 assimilation and transpiration of a Panamanian tropical forest. *New Phytol.*, **238**, 2345–2362, https://doi.org/10.1111/nph.18901.

Langford, Z. L., J. Kumar, and F. M. Hoffman, 2017: Convolutional neural network approach for mapping Arctic vegetation using multi-sensor remote sensing fusion. *Proc. 2017 IEEE Int. Conf. on Data Mining Workshops*, New Orleans, LA, IEEE, 322–331, https://doi.org/10.1109/ICDMW.2017.48.

——, ——, ——, A. L. Breen, and C. M. Iversen, 2019: Arctic vegetation mapping using unsupervised training datasets and convolutional neural networks. *Remote Sens.*, **11**, 69, https://doi.org/10.3390/rs11010069.

Lawrence, D. M., and Coauthors, 2019: The Community Land Model version 5: Description of new features, benchmarking, and impact of forcing uncertainty. *J. Adv. Model. Earth Syst.*, **11**, 4245–4287, https://doi.org/10.1029/2018MS001583.

Lin, L., X. Wei, P. Luo, S. Wang, D. Kong, and J. Yang, 2023: Ecological security patterns at different spatial scales on the Loess Plateau. *Remote Sens.*, **15**, 1011, https://doi.org/10.3390/rs15041011.

Liu, S., and Coauthors, 2011: Simulating the impacts of disturbances on forest carbon cycling in North America: Processes, data, models, and challenges. *J. Geophys. Res.*, **116**, G00K08, https://doi.org/10.1029/2010JG001585.

Maclean, I. M. D., 2020: Predicting future climate at high spatial and temporal resolution. *Global Change Biol.*, **26**, 1003–1011, https://doi.org/10.1111/gcb.14876.

Maneta, M. P., and N. L. Silverman, 2013: A spatially distributed model to simulate water, energy, and vegetation dynamics using information from regional climate models. *Earth Interact.*, **17**, https://doi.org/10.1175/2012EI000472.1.

Mao, J., Y. Wang, D. Ricciuto, S. Mahajan, F. Hoffman, X. Shi, and G. Prakash, 2021: AI-based integrated modeling and observational framework for improving seasonal to decadal prediction of terrestrial ecohydrological extremes. Tech. Rep. AI4ESP-1089, 5 pp., https://doi.org/10.2172/1769666.

Massoud, E. C., 2019: Emulation of environmental models using polynomial chaos expansion. *Environ. Modell. Software*, **111**, 421–431, https://doi.org/10.1016/j.envsoft.2018.10.008.

——, A. J. Purdy, B. O. Christoffersen, L. S. Santiago, and C. Xu, 2019: Bayesian inference of hydraulic properties in and around a white fir using a process-based ecohydrologic model. *Environ. Modell. Software*, **115**, 76–85, https://doi.org/10.1016/j.envsoft.2019.01.022.

——, A. A. Bloom, M. Longo, J. T. Reager, P. A. Levine, and J. R. Worden, 2022: Information content of soil hydrology in a west Amazon watershed as informed by GRACE. *Hydrol. Earth Syst. Sci.*, **26**, 1407–1423, https://doi.org/10.5194/hess-26-1407-2022.

Miao, C., H. Ashouri, K.-L. Hsu, S. Sorooshian, and Q. Duan, 2015: Evaluation of the PERSIANN-CDR daily rainfall estimates in capturing the behavior of extreme precipitation events over China. *J. Hydrometeor.*, **16**, 1387–1396, https://doi.org/10.1175/JHM-D-14-0174.1.

Mishra, U., S. Gautam, W. J. Riley, and F. M. Hoffman, 2020: Ensemble machine learning approach improves predicted spatial variation of surface soil organic carbon stocks in data-limited northern circumpolar region. *Front. Big Data*, **3**, 528441, https://doi.org/10.3389/fdata.2020.528441.

——, K. Yeo, K. Adhikari, W. J. Riley, F. M. Hoffman, C. Hudson, and S. Gautam, 2022: Empirical relationships between environmental factors and soil organic carbon produce comparable prediction accuracy as the machine learning. *Soil. Sci. Soc. Amer. J.*, **86**, 1611–1624, https://doi.org/10.1002/saj2.20453.

Nativi, S., P. Mazzetti, and M. Craglia, 2021: Digital ecosystems for developing digital twins of the Earth: The destination Earth case. *Remote Sens.*, **13**, 2119, https://doi.org/10.3390/rs13112119.

Peters, J., B. De Baets, N. E. Verhoest, R. Samson, S. Degroeve, P. De Becker, and W. Huybrechts, 2007: Random forests as a tool for ecohydrological distribution modelling. *Ecol. Modell.*, **207**, 304–318, https://doi.org/10.1016/j.ecolmodel.2007.05.011.

Potapov, P., and Coauthors, 2021: Mapping global forest canopy height through integration of GEDI and Landsat data. *Remote Sens. Environ.*, **253**, 112165, https://doi.org/10.1016/j.rse.2020.112165.

Rasp, S., M. S. Pritchard, and P. Gentine, 2018: Deep learning to represent subgrid processes in climate models. *Proc. Natl. Acad. Sci. USA*, **115**, 9684–9689, https://doi.org/10.1073/pnas.1810286115.

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, 2019: Deep learning and process understanding for data-driven Earth system science. *Nature*, **566**, 195–204, https://doi.org/10.1038/s41586-019-0912-1.

Rings, J., T. Kamai, M. Kandelous, P. Hartsough, J. Simunek, J. A. Vrugt, and J. W. Hopmans, 2013: Bayesian inference of tree water relations using a soil-tree-atmosphere continuum model. *Procedia Environ. Sci.*, **19**, 26–36, https://doi.org/10.1016/j.proenv.2013.06.004.

Rodriguez-Iturbe, I., 2000: Ecohydrology: A hydrologic perspective of climate-soil-vegetation dynamics. *Water Resour. Res.*, **36**, 3–9, https://doi.org/10.1029/1999WR900210.

Rolnick, D., and Coauthors, 2022: Tackling climate change with machine learning. *ACM Comput. Surv.*, **55**, 42, https://doi.org/10.1145/3485128.

Russakovsky, O., and Coauthors, 2015: ImageNet large scale visual recognition challenge. *Int. J. Comput. Vision*, **115**, 211–252, https://doi.org/10.1007/s11263-015-0816-y.

Santi, E., S. Paloscia, S. Pettinato, and G. Fontanelli, 2016: Application of artificial neural networks for the soil moisture retrieval from active and passive microwave spaceborne sensors. *Int. J. Appl. Earth Obs. Geoinf.*, **48**, 61–73, https://doi.org/10.1016/j.jag.2015.08.002.

Saunders, A., D. M. Drew, and W. Brink, 2021: Machine learning models perform better than traditional empirical models for stomatal conductance when applied to multiple tree species across different forest biomes. *Trees For. People*, **6**, 100139, https://doi.org/10.1016/j.tfp.2021.100139.

Schneider, T., S. Lan, A. Stuart, and J. Teixeira, 2017: Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophys. Res. Lett.*, **44**, 12 396–12 417, https://doi.org/10.1002/2017GL076101.

Scott, R. L., T. E. Huxman, D. G. Williams, and D. C. Goodrich, 2006: Ecohydrological impacts of woody-plant encroachment: Seasonal patterns of water and carbon dioxide exchange within a semiarid riparian environment. *Global Change Biol.*, **12**, 311–324, https://doi.org/10.1111/j.1365-2486.2005.01093.x.

Seidl, R., and Coauthors, 2011: Modelling natural disturbances in forest ecosystems: A review. *Ecol. Modell.*, **222**, 903–924, https://doi.org/10.1016/j.ecolmodel.2010.09.040.

Shivaprakash, K. N., N. Swami, S. Mysorekar, R. Arora, A. Gangadharan, K. Vohra, M. Jadeyegowda, and J. M. Kiesecker, 2022: Potential for artificial intelligence (AI) and machine learning (ML) applications in biodiversity conservation, managing forests, and related services in India. *Sustainability*, **14**, 7154, https://doi.org/10.3390/su14127154.

Slater, L., and Coauthors, 2023: Hybrid forecasting: Blending climate predictions with AI models. *Hydrol. Earth Syst. Sci.*, **27**, 1865–1889, https://doi.org/10.5194/hess-27-1865-2023.

Sprenger, M., S. Sheila, J. P. Dennedy-Frank, and E. Woodburn, 2021: Preferential flow in subsurface hydrology: From a century of denial to a decade of addressing it via ML? Tech. Rep. AI4ESP1125, 5 pp., https://doi.org/10.2172/1769765.

Steidinger, B. S., and Coauthors, 2019: Climatic controls of decomposition drive the global biogeography of forest-tree symbioses. *Nature*, **569**, 404–408, https://doi.org/10.1038/s41586-019-1128-0.

Sukanya, S., and S. Joseph, 2023: Climate change impacts on water resources: An overview. *Visualization Techniques for Climate Change with Machine Learning and Artificial Intelligence*, A. Srivastav et al., Eds., Elsevier, 55–76, https://doi.org/10.1016/B978-0-323-99714-0.00008-X.

Tague, C., and J. Frew, 2021: Visualization and ecohydrologic models: Opening the box. *Hydrol. Processes*, **35**, e13991, https://doi.org/10.1002/hyp.13991.

Tao, Y., X. Gao, K. Hsu, S. Sorooshian, and A. Ihler, 2016: A deep neural network modeling framework to reduce bias in satellite precipitation products. *J. Hydrometeor.*, **17**, 931–945, https://doi.org/10.1175/JHM-D-15-0075.1.

Tsai, W.-P., D. Feng, M. Pan, H. Beck, K. Lawson, Y. Yang, J. Liu, and C. Shen, 2021: From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nat. Commun.*, **12**, 5988, https://doi.org/10.1038/s41467-021-26107-z.

Vico, G., and Coauthors, 2015: Climatic, ecophysiological, and phenological controls on plant ecohydrological strategies in seasonally dry ecosystems. *Ecohydrology*, **8**, 660–681, https://doi.org/10.1002/eco.1533.

Vitharana, U. W. A., U. Mishra, J. D. Jastrow, R. Matamala, and Z. Fan, 2017: Observational needs for estimating Alaskan soil carbon stocks under current and future climate. *J. Geophys. Res. Biogeosci.*, **122**, 415–429, https://doi.org/10.1002/2016JG003421.

Wan, T., B. H. Covert, C. N. Kroll, and C. R. Ferguson, 2022: An assessment of the national water model's ability to reproduce drought series in the northeastern United States. *J. Hydrometeor.*, **23**, 1929–1943, https://doi.org/10.1175/JHM-D-21-0226.1.

Wang, C., B. Fu, L. Zhang, and Z. Xu, 2019: Soil moisture–plant interactions: An ecohydrological review. *J. Soils Sediments*, **19**, 1–9, https://doi.org/10.1007/s11368-018-2167-0.

Wang, L., S. Manzoni, S. Ravi, D. Riveros-Iregui, and K. Caylor, 2015: Dynamic interactions of ecohydrological and biogeochemical processes in water-limited systems. *Ecosphere*, **6**, 133, https://doi.org/10.1890/ES15-00122.1.

Wang, Y., P. Köhler, L. He, R. Doughty, R. K. Braghiere, J. D. Wood, and C. Frankenberg, 2021a: Testing stomatal models at the stand level in deciduous angiosperm and evergreen gymnosperm forests using CliMA land (v0.1). *Geosci. Model Dev.*, **14**, 6741–6763, https://doi.org/10.5194/gmd-14-6741-2021.

——, J. Mao, M. Jin, F. M. Hoffman, X. Shi, S. D. Wullschleger, and Y. Dai, 2021b: Development of observation-based global multilayer soil moisture products for 1970 to 2016. *Earth Syst. Sci. Data*, **13**, 4385–4405, https://doi.org/10.5194/essd-13-4385-2021.

Williams, J. W., S. T. Jackson, and J. E. Kutzbach, 2007: Projected distributions of novel and disappearing climates by 2100 AD. *Proc. Natl. Acad. Sci. USA*, **104**, 5738–5742, https://doi.org/10.1073/pnas.0606292104.

Xu, X., and A. T. Trugman, 2021: Trait-based modeling of terrestrial ecosystems: Advances and challenges under global change. *Curr. Climate Change Rep.*, **7** (1), 1–13, https://doi.org/10.1007/s40641-020-00168-6.

Yan, H., and Coauthors, 2023: Characterizing uncertainty in community land model version 5 hydrological applications in the United States. *Sci. Data*, **10**, 187, https://doi.org/10.1038/s41597-023-02049-7.

Yang, X., and Coauthors, 2023: Leaf angle as a leaf and canopy trait: Rejuvenating its role in ecology with new technology. *Ecol. Lett.*, **26**, 1005–1020, https://doi.org/10.1111/ele.14215.

Yang, Y., and Coauthors, 2022: CARDAMOM-FluxVal version 1.0: A FLUXNET-based validation system for CARDAMOM carbon and water flux estimates. *Geosci. Model Dev.*, **15**, 1789–1802, https://doi.org/10.5194/gmd-15-1789-2022.

Yu, Y., and Coauthors, 2022: Machine learning–based observation-constrained projections reveal elevated global socioeconomic risks from wildfire. *Nat. Commun.*, **13**, 1250, https://doi.org/10.1038/s41467-022-28853-0.

Zhang, K., G. Zhu, N. Ma, H. Chen, and S. Shang, 2022: Improvement of evapotranspiration simulation in a physically based ecohydrological model for the groundwater–soil–plant–atmosphere continuum. *J. Hydrol.*, **613**, 128440, https://doi.org/10.1016/j.jhydrol.2022.128440.