Using AI to build a hydrobiogeochemical soil model

Authors

Beth Drewniak, Julie Jastrow, Richard Tran Mills, Roser Matamala, Kenneth Kemner, Alexandre Renchon, Miquel Gonzalez-Meler, Pamela Weisenhorn, Zhengchun Liu, Julie Bessac, Kaiyu Guan, Bin Peng, Hanghang Tong, Andrew Margenot, Kathe Todd-Brown

Focal Area(s)

This whitepaper responds to focal areas 2 and 3 and intends to integrate complex observational data collected across the land-atmosphere domain to inform models and explore the role of soil as an integrator of land-atmosphere hydrobiogeochemical cycles.

Science Challenge

Soil water content is a function of inputs from precipitation and outputs via evaporation, transpiration, lateral flow, and vertical percolation, and is sensitive to biogeochemical processes. As such, soils serve as an ideal integrator of atmospheric, hydrological, and biogeochemical processes affecting the water cycle. In addition, soil water retention capacity, infiltration rates, and hydraulic conductivity can buffer or exacerbate the effects of extreme precipitation events (e.g., flooding, runoff, subsurface transport, erosion, greenhouse gas emissions) and mitigate the impact of droughts and heat waves on land systems (e.g., fire, crop failure). However, integrating water cycle measurements spanning different land-atmosphere compartments across scales is a fundamental barrier for numerical model predictability. A significant challenge is that each domain (soil, hydrology, biology, and atmosphere) typically collects different sets of data at different temporal and spatial frequencies/scales, and even different dimensionalities (2D vs 3D). To implement soil as an integrator of the water cycle in land models, we suggest that novel machine learning (ML) tools can be developed to effectively simulate complex landscapes across various domains and scales, extended to regions with sparse or no data. The ultimate goals are to improve predictive understanding of land-atmosphere interactions and to extend the predictability of current Earth System Models (ESMs) through better integration of hydrological and biogeochemical data. We envision a framework in which: (1) ML-aided data reconstructions enable the merger of data sources into a unified geospatial product; (2) automated detection techniques are used to improve the knowledge of complex soil processes and interactions; and (3) this knowledge is leveraged and incorporated into models through AI-based emulators to distinctly connect the land and atmospheric compartments of the water cycle in models.

Rationale

Soils are heterogeneous and complex at multiple scales. Variations in soil hydrological processes are inherently linked to properties such as particle size distribution, porosity, permeability, clay mineralogy, depth, and organic matter content that vary across landscapes and regions. Thus, soil hydrobiogeochemical functions and their interconnections occur at multiple spatiotemporal scales across the landscape and with depth. In general, this heterogeneity is not well represented in models, particularly ESMs designed to address global-century level climate questions. For example, small variations in topography can lead to differences in ecosystem function via soil moisture, microbial activity, biogeochemical reactions, and gas emissions. Data sets are often collected at field or plot scales (which are coarser than the scales at which underlying processes occur) and then used to empirically parameterize components of land surface models (LSMs) with little understanding of possible feedbacks to the atmosphere. As LSMs become more complex and better coupled to atmospheric models, it will become necessary to find solutions for managing the dimensionality created by numerous interacting

DOI: https://doi.org/10.2172/1769793
DOE BER, AI4ESP 2021 #AI4ESP1035 (https://www.ai4esp.org)

feedback systems and the heterogeneity of land processes. Upscaling of heterogeneous land processes is a well-recognized problem in ESMs. These representations may reproduce reasonable behavior over long time scales on an average model grid cell but are missing the emergent properties that ultimately underpin soil hydrological integration of air, water, land and biology. For this reason, extrapolating to the future under changing conditions is also likely to miss key responses in ecosystems. By explicitly exploring these scaling issues in the context of land-atmosphere multi-scaled data sets, we can rapidly advance the field. Furthermore, by using soil as an integrator of the atmospheric-land water cycle we can elucidate how soil processes, properties, and pre-existing conditions affect ecosystem responses to extreme events and, thereby, improve their predictability.

Narrative

Similar to the move towards trait-based representations of vegetation in modern LSMs, we propose that water-cycle predictability can be greatly improved by incorporating soil traits (properties) that represent key soil functions [e.g., 1-5] affecting land hydrobiogeochemical processes and land-atmosphere interconnections in ESMs. Because soil properties can be predicted from multivariate relationships with environmental data that serve as proxies for soil formation and evolution, the spatial distributions of these covariates can be used to generate spatially explicit predictions of soil traits from observational datasets. Traditionally, digital representations of soil property distributions have been generated by using linear regression, kriging, and hybrid approaches [6-9], but increasingly the utility of various ML and deep learning approaches are being explored [10-14]. Recent national and international efforts to build and harmonize soil information systems are accelerating due to expansion of traditional data streams, creation of standardized databases, advancements in proximal soil sensing, and efforts to develop new *in situ* sensors and sensor networks [15-23]. In addition, high-resolution satellite remote sensing of surface properties (e.g., land use/land cover types, elevation), states (such as leaf area index, biomass, surface soil moisture), and energy, water and carbon fluxes (such as evapotranspiration and gross primary production) are becoming increasingly available. These satellite-based observations have global coverage with high spatiotemporal resolutions, providing new or enhanced environmental predictor data streams [24-29]. Thus, we believe the time is right for coupling AI with these exploding data sources to build integrated soil-aware land-atmospheric domain datasets that can both improve ESMs and inform development/deployment of targeted observation networks to reduce uncertainties and increase predictive understanding.

Data Reconstruction*:* AI can be used to integrate, unify, and harmonize data that expand across scales. This requires integration of a hierarchy of observational datasets that range from high to low frequency (e.g., hourly, daily, one-time measurement, etc.), have been collected across different time scales (e.g., once, seasonal, annual, decadal), and are distributed with varied spatial intensity across different regions. We propose to use AI-based multiscale techniques to reconstruct (interpolate) and link data across scales, for example mapping multiple irregular sensor outputs onto regular grids. Soils in general and soil water in particular show a multitude of interactions and associations across various scales, and these multiscale features can be embedded into the reconstruction scheme. For example, an open challenge in the field is linking high-resolution remote sensing products and digital maps of soil properties with dynamic field-scale observations (e.g., soil moisture sensor networks, runoff monitoring, streamflow, eddy covariance data). AI has several methods that are well suited to this challenge. Artificial neural networks have proven successful at embedding multiple sources of data [30,31] and mapping complex non-linear input-outputs relationships. In particular, techniques such as generative adversarial networks have been successful at enhancing the resolution of outputs and will be used to obtain the most detailed representation of the quantities of interest [32]. In addition, Gaussian processes are known for

accommodating unstructured data, interpolating in a multidimensional fashion (space-time, scale-aware, multiple variables, etc.) and for providing uncertainty associated with such reconstructions. The use of Gaussian process techniques provides uncertainty measures and can be leveraged with environmental information to create an optimization framework to guide the placement of new sensors.

Association Detection and Automation: Soil science has a long history of interoperating data from other measurements (e.g., pedotransfer functions [33]) and working with physics-based simulations to examine associations between data. These models, however, typically rely on sparse data and/or 'expert tuned' models. We can leverage new AI methods to launch the field forward by using clustering and segmentation techniques to detect and extract associations between variables across high volumes of data and hierarchies of scale. For instance, artificial neural networks (e.g., Autoencoder) can be trained to extract the most representative features from the high dimensional reconstructed observations. Then visualization algorithms such as t-distributed Stochastic Neighbor Embedding can be used for mapping the compressed features to a 2D plane for exploring similarity and associations. Furthermore, semi-supervised extensions can be built to include existing knowledge, such as extracted spectral coherence for sparse data. Creation of reconstructed datasets with the same dimensions in time and space, will enable mathematical analysis of coherence (from 0 to 1) at different scales by using, e.g., wavelet analysis. For example, we expect high coherence between photosynthesis and radiation at hourly and daily time scales in vegetated areas. At inter-annual scales, some areas will have high coherence between photosynthesis and radiation, or precipitation, or temperature. Similarly, in space, soil CO_2 efflux might have high coherence with soil moisture at the centimeter scale, whereas it might have high coherence with vegetation density at kilometer scales. Such mapping of coherence between variables in time and space at various scales would be insightful for many domains of science and would inform areas that need more research to understand the underlying mechanisms of apparent system behaviors. Furthermore, the exact same analysis can be done with ESM variables to enable comparisons with empirical data and further inform on mismatch and agreement with the data — thereby guiding our understanding of emergent properties and providing a new approach to benchmark ESMs at various scales. Finally, these insights will be used to understand which soil properties and functions are dominant controllers/predictors at different scales. Knowledge gained from AI can incorporate heterogeneity in models by informing which processes to include in sub-grid or refined grids in ESMs.

Emulators: Finally, models can be improved by the insight gained from the development of emulators to capture emergent properties. AI-based emulators could be designed for scale-aware hydrobiogeochemical models as a 3D controller or strategy for representing fine scale soil processes in ESMs by using approaches such as the deep neural architecture search, which requires limited training data [34]. Learning from fine-scale modeling over targeted regions or environmental conditions is also a promising pathway forward. The ESM community can benefit from recent advances in fine-scale modeling (such as 3D simulation of coupled surface-subsurface water and nutrient flows). With the help of advanced AI/ML techniques, the most relevant features controlling a specific process can be identified within the simulated database. Similarly, emulators can also be built by learning from the simulated database, especially when the model simulations are constrained by observations. In addition, recent emergence of physics-guided machine learning provides another opportunity for learning from both model simulations and observations, which usually can lead to superior performance compared to learning from either model simulations or observations alone. Finally, embedding fine scale models within the larger modeling framework at representative locations via a clustering approach (similar to superparameterization approaches in the atmosphere) can be used to scale up high resolution processes to a greater region. AI techniques can identify the locations of representative clusters and parameterize the fine scale models.

Using AI to build a hydrobiogeochemical soil model

Suggested Partners/Experts (Optional)

1. International Soil Modeling Consortium (ISMC)
2. Earth Science Information Partners (ESIP)
3. Drought-Net
4. Natural Resources Conservation Service (NRCS) and Kellogg Soil Survey Laboratory (KSSL)
5. National Ecological Observatory Network (NEON)

References (Optional)

1. Arrouays, D., *et al*. 2014. GlobalSoilMap: Toward a fine-resolution global grid of soil properties. *Advances in Agronomy*, 125:93-134.
2. Ross, C.W., L. Prihodko, J. Anchang, S. Kumar, W. Ji, and N.P. Hanan. 2018. HYSOGs250m, global gridded hydrologic soil groups for curve-number-based runoff modeling. *Scientific Data*, 5:150091.
3. Chaney, N.W., *et al*. 2019. POLARIS soil properties: 30-m probabilistic maps of soil properties over the contiguous United States. *Water Resources Research*, 55:2916-2938.
4. Chen, S., *et al*. 2019. National estimation of soil organic carbon storage potential for arable soils: A data-driven approach coupled with carbon-landscape zones. *Science of the Total Environment*, 666:355-367.
5. Arrouays, D., L. Poggio, O.A.S. Guerrero, and V.L. Mulder. 2020. Digital soil mapping and GlobalSoilMap. Main advances and ways forward. *Geoderma Regional*, 21:e00265.
6. Odeh, I.O.A., A.B. McBratney, and D.J. Chittleborough. 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. *Geoderma*, 63:197-214.
7. McBratney, A.B., I.O.A., Odeh, T.F., Bishop, M.S. Dunbar, and T.M. Shatar. 2000. An overview of pedometric techniques for use in soil survey. *Geoderma*, 97:293-327.
8. Hengl, T., G.B. Heuvelink, and A. Stein. 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma*, 120:75-93.
9. Hengl, T., G.B. Heuvelink, and D.G. Rossiter, 2007. About regression-kriging: From equations to case studies. *Computers & Geosciences*, 33:1301-1315.
10. Hengl, T., *et al*. 2017. SoilGrids250m: Global gridded soil information based on machine learning. *PLoS ONE*, 12:e0169748.
11. Padarian, J., B. Minasny, and A.B. McBratney. 2019. Using deep learning for digital soil mapping. *SOIL*, 5:79-89.
12. Wadoux, A.M.C., J. Padarian, and B. Minasny. 2019. Multi-source data integration for soil mapping using deep learning. *SOIL*, 5:107-119.
13. Padarian, J., B. Minasny, and A.B. McBratney. 2020. Machine learning and soil sciences: A review aided by machine learning tools. *SOIL*, 6:35-52.
14. Wadoux, A.M.J.-C., M. Román-Dobarco, and A.B. McBratney. 2020. Perspectives on data-driven soil research. *European Journal of Soil Science*. (in press) https://doi.org/10.1111/ejss.13071
15. Viscarra-Rossel, R.A., V.I. Adamchuk, K.A. Sudduth, N.J. McKenzie, and C. Lobsey. 2011. Proximal soil sensing: An effective approach for soil measurements in space and time. *Advances in Agronomy*, 113:243-291.
16. Hartemink, A.E., and B. Minasny. 2014. Towards digital soil morphometrics. *Geoderma*, 230:305-317.
17. Viscarra-Rossel, R.A., *et al*. 2016. A global spectral library to characterize the world's soil. *Earth-Science Reviews*, 155:198-230.
18. Dong, J., and T.E. Ochsner, 2018. Soil texture often exerts a stronger influence than precipitation on mesoscale soil moisture patterns. *Water Resources Research*, 54:2199-2211.
19. Batjes, N.H., E. Ribeiro, and A. van Oostrum. 2020. Standardized soil profile data to support global mapping and modelling (WoSIS snapshot 2019). *Earth System Science Data*, 12:299–320.

20. Bond-Lamberty, B., *et al*. 2020. COSORE: A community database for continuous soil respiration and other soil-atmosphere greenhouse gas flux data. *Global Change Biology*, 26:7268-7283.
21. Zhang Y., A.E. Hartemink, J. Huang, and P.A. Townsend. 2021. Synergistic use of hyperspectral imagery, Sentinel-1 and LiDAR improves mapping of soil physical and geochemical properties at the farm-scale. *European Journal of Soil Science*. (in press) https://doi.org/10.1111/ejss.13086
22. Dorigo, W., *et al*. 2021. The International Soil Moisture Network: serving Earth system science for over a decade. *Hydrology & Earth System Sciences Discussions*. [preprint], https://doi.org/10.5194/hess-2021-2, in review.
23. FAO Global Soil Partnership. Pillar 4: Information and Data. http://www.fao.org/global-soil-partnership/pillars-action/4-information-data/en/
24. Poggio, L., A. Gimona, and M.J. Brewer. 2013. Regional scale mapping of soil properties and their uncertainty with a large number of satellite-derived covariates. *Geoderma*, 209:1-14.
25. Demattê, J.A.M., C.T. Fongaro, R. Rizzo, and J.L. Safanelli. 2018. Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images. *Remote Sensing of Environment*, 212:161-175.
26. Artz, R.R., *et al*. 2019. The potential for modelling peatland habitat condition in Scotland using long-term MODIS data. *Science of the Total Environment*, 660:429-442.
27. Castaldi, F., *et al*. 2019. Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands. *ISPRS Journal of Photogrammetry and Remote Sensing*, 147:267-282.
28. Loiseau, T., *et al*. 2019. Satellite data integration for soil clay content modelling at a national scale. *International Journal of Applied Earth Observation and Geoinformation*, 82:101905.
29. Vaudour, E., C. Gomez, Y. Fouad, and P. Lagacherie. 2019. Sentinel-2 image capacities to predict common topsoil properties of temperate and Mediterranean agroecosystems. *Remote Sensing of Environment*, 223:21-33.
30. Wang, J., Z. Liu, I. Foster, W. Chang, R. Kettimuthu, and V.R. Kotamarthi. 2021. Fast and accurate learned multiresolution dynamical downscaling for precipitation. *Geoscientific Model Development Discussions*. [preprint] https://doi.org/10.5194/gmd-2020-412, in review.
31. Chai, X., *et al*. 2020. Deep learning for irregularly and regularly missing data reconstruction. *Scientific Reports*, 10:3302.
32. Azevedo, L., G. Paneiro, A. Santos, and A. Soares. 2020. Generative adversarial network as a stochastic subsurface model reconstruction. *Computational Geosciences*, 24:1673-1692.
33. Van Looy, K., *et al*. 2017. Pedotransfer functions in Earth system science: challenges and perspectives. *Reviews of Geophysics*, 55:1199-1256.
34. Kasim, M.F., *et al*. 2020. Building high accuracy emulators for scientific simulations with deep neural architecture search. arXiv:2001.08055v2. https://arxiv.org/pdf/2001.08055.pdf