Computationally Tractable High-Fidelity Representation of Global Hydrology in ESMs via Machine Learning Approaches to Scale-Bridging

Richard Tran Mills[1] (rtmills@anl.gov), Forrest M. Hoffman[2], Jitendra Kumar[2], Robert Jacob[1], Zachary Langford[2], Sarat Sreepathi[2], Nathan Collier[2]

Focal Areas: This paper responds primarily to Focal Area 2, focusing on AI techniques to improve model fidelity.

Science Challenge: "Hyperresolution" [1, 2] land surface models (LSMs) running at far higher resolution than typically employed in global Earth system models (ESMs) can help answer critical questions about the water cycle and associated ecosystem and biogeochemical feedbacks. Even with all foreseeable advances in computing power and efficient solver algorithms, however, employing hyperresolution LSMs inside ESMs for studies of long-term global climate is not computationally feasible. Instead, we argue for incorporating the fidelity of hyperresolution LSMs only where and when it is needed by using machine learning approaches to scale-bridging.

Rationale: Typical ESM resolutions are too coarse to represent important nonlinear hydrologic responses driven by complex spatial heterogeneity; this becomes particularly important in studying hydrologic extremes, where negative and positive anomalies may be simultaneously discernible in a catchment at high resolution, but this internal variability is smoothed out at coarser resolutions. An obvious example calling for hyperresolution is flood modeling, where resolving complex, fine-scale patterns of inundation can be important to capture biogeochemical drivers of the climate system. In the Amazon Basin, for example, floods during the rainy season control significant CO_2 outgassing (on the order of 0.5 Gt per year of carbon [3]), much of it from small channels [4] that cannot be represented at coarse resolution.

Periglacial regions in the Arctic, where local hydrology is being strongly perturbed by rapid permafrost degradation due to rapid warming, illustrate the challenges in upscaling the effects of fine-scale land-surface processes to larger scales due to their complex multi-scale organization, in which extreme heterogeneity can be observed at typical climate model cell resolution (100 km) down to sub-meter resolution (Figure 1(a)). To develop a strategy for systematically collecting field measurements relevant to constraining models in this region, in [5] we applied an unsupervised learning approach to downscaled general circulation model results and observational data for the State of Alaska to define a set of eco-climatic regions with relatively homogeneous attributes, at multiple levels of division across two decadal time periods, and examined how their distributions are expected to shift across Alaska due to projected climate change. Using a metric of "representativeness" based on Euclidean distances in the n-dimensional state space employed in our cluster analysis, optimal sampling locations (the "realized centroids") that are most representative of the conditions in each eco-climatic region were obtained, and candidate locations for measurement sites were identified (see Figure 1(c)). The cluster analysis and data space regressions offer a ba-

[1]Argonne National Laboratory, Lemont, IL 60439
[2]Oak Ridge National Laboratory, Oak Ridge, TN 37831

DOI: https://doi.org/10.2172/1769690

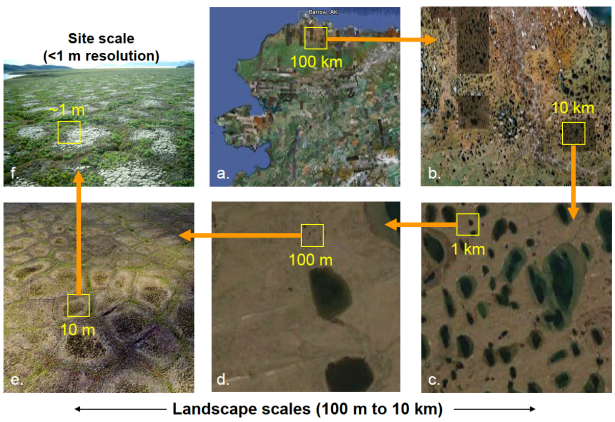
sis for upscaling and extrapolating measurements to land areas within and beyond the sampling domains, and form an important part of our approach.

Narrative: To circumvent the tremendous computational expense of fully-resolved ultra-high resolution models, we propose an approach inspired partly by the "super-parameterization" [6, 7] approaches developed by the atmospheric modeling community. These avoid the tremendous computational expense of cloud system-resolving models while still realistically representing small-scale and mesoscale processes, by embedding a small-scale 2D or quasi-3D model inside each large-scale (here denoted "global scale" for convenience) cell. The embedded small-scale models generally employ periodic lateral boundary conditions, and do not interact with each other except indirectly through the fluxes on the global-scale grid. This approach essentially consists of two models with two sets of model variables: the large-scale models and variables of the coarse, global grid, and the fine-scale models and variables of the embedded, small-scale models. The two sets of model variables are coupled by enforcing the property that the horizontal average of a small-scale variable is exactly equal to the value of the corresponding large-scale value.

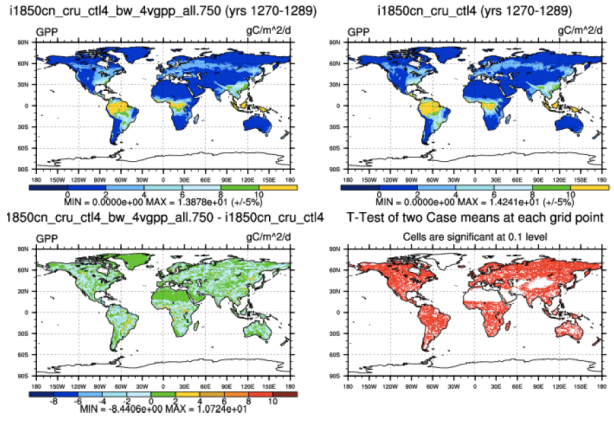
Although 2D models are attractive from the standpoint of computational expense, in most landscapes the hydrologic processes require full 3D representation. Embedding finely resolved 3D small-scale grids in every global-scale grid is not computationally feasible, however, even if the individual small-scale grids do not communicate with each other. However, *we believe that unsupervised (or semi-supervised) techniques similar to those we have applied to choosing measurement sites in ecological sampling networks can be used to choose a sparse set of "measurement" sites where fine-scale, 3D models are to be embedded*. A basic idea is to periodically run an on-line clustering using the parameters and state variables from all cells in the global grid (perhaps with ancillary fine-scale observational data); group cells with similar states, properties, and forcings together into clusters; choose a representative subset of cells from each cluster, and run an embedded fine-scale model—coupled to the global scale in a manner very similar to that used in atmospheric super-parameterization models. Quantities of interest, e.g., biogeochemical reaction rates, from the sparse collection of fine-scale models are then mapped back to the other global cells that are members of the same (or similar) clusters that do not have an embedded fine-scale model. The simplest approach is to "paint by numbers": if only one member of a cluster has an embedded model, assign the same upscaled value from that member to all others of the cluster; if there are multiple members of the cluster with embedded fine-scale models runs, assign other members the value by some intra- or inter-cluster regression. Such an approach might work surprisingly well: geospatiotemporal clustering has been used to dramatically reduce the dimensionality of global CLM simulation output [8] while maintaining high accuracy (Figure 1(b)).

The naive approach outlined above has several potential issues: Some attempt should be made to re-use previous fine-scale runs with input parameters similar to those in a current cluster. Very high-dimensional features may be needed, which may make obtaining good clusterings difficult (approaches such as clustering within a latent feature space determined by a deep autoencoder neural network [9] may help). The level of division k needed may be very high; furthermore, it is not clear that all clusters will represent regions for which fine-scale models are needed. It may be better to approach things from the perspective of training of a non-parametric regressor

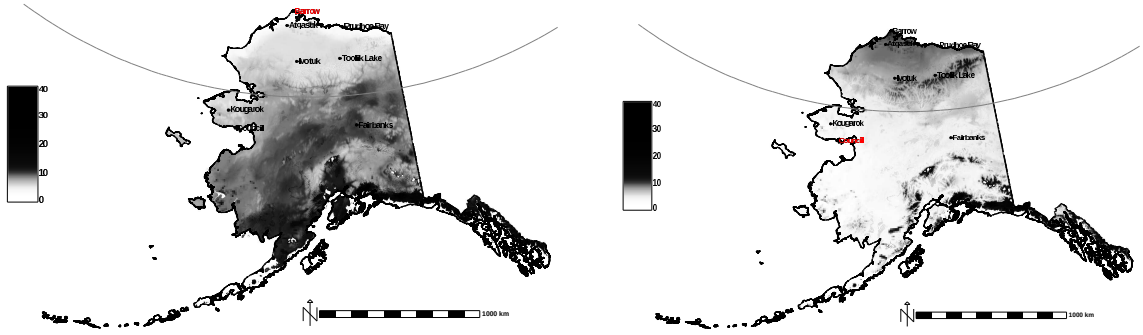
(our "process emulator") using ideas from active learning and optimal experimental design [10], in which new training points (fine-scale model runs, in our case) are chosen according to what will best reduce some measure of uncertainty or maximize information gain. This needs to be an online process, due to the very large state space of the global model and because rapid climate change is pushing the hydrologic cycle far from stationarity. Adequately exploring this topic will require an ambitious, sustained research program involving members of the climate science, computational mathematics, and machine learning communities, and putting it into practice will require overcoming substantial software engineering and model coupling challenges. We believe that the potential payoff in terms of dramatically increased fidelity in long-term modeling of global hydrology, however, make this a worthwhile investment.



(a) Permafrost region scaling



(b) Clustering-based dimensionality reduction



(c) Alaska representativeness

Figure 1: Different aspects of spatial scaling and representativeness in landscapes. a) A zoom-in over patterned ground on the North Slope of Alaska, near Barrow, AK, illustrates distinct (and geometric) small-scale features. Figure courtesy of Peter Thornton, ORNL. b) A comparison of Gross Primary Productivity (GPP) from an approximately 60,000 cell CLM simulation (top right) and from a reduced version (top left) in which geospatiotemporal clustering has been used to identify 750 clusters; GPP values on the top left map have been filled in by assigning the GPP value associated with the centroid of a cluster to all cells that are members of the cluster. Figure courtesy of J. Kumar and F. M. Hoffman of ORNL, from unpublished work. c) A comparison of representativeness (closeness in terms of Euclidean distance in the data-space of eco-climatic variables used in the cluster analysis) for present-day conditions for two sites considered in [5]. White to light gray areas are well-represented by the site; dark gray to black areas are poorly represented.

References

[1] E. F. Wood, J. K. Roundy, T. J. Troy, L. P. H. van Beek, M. F. P. Bierkens, E. Blyth, A. de Roo, P. Döll, M. Ek, J. Famiglietti, D. Gochis, N. van de Giesen, P. Houser, P. R. Jaffé, S. Kollet, B. Lehner, D. P. Lettenmaier, C. Peters-Lidard, M. Sivapalan, J. Sheffield, A. Wade, and P. Whitehead, "Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring earth's terrestrial water," *Water Resources Research*, vol. 47, no. 5, pp. n/a–n/a, 2011. [Online]. Available: http://dx.doi.org/10.1029/2010WR010090

[2] M. F. P. Bierkens, V. A. Bell, P. Burek, N. Chaney, L. E. Condon, C. H. David, A. de Roo, P. Döll, N. Drost, J. S. Famiglietti, M. Flörke, D. J. Gochis, P. Houser, R. Hut, J. Keune, S. Kollet, R. M. Maxwell, J. T. Reager, L. Samaniego, E. Sudicky, E. H. Sutanudjaja, N. van de Giesen, H. Winsemius, and E. F. Wood, "Hyper-resolution global hydrological modelling: what is next?" *Hydrological Processes*, vol. 29, no. 2, pp. 310–320, 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.10391

[3] J. E. Richey, J. M. Melack, A. K. Aufdenkampe, V. M. Ballester, and L. L. Hess, "Outgassing from amazonian rivers and wetlands as a large tropical source of atmospheric CO2," *Nature*, vol. 416, pp. 617–620, 2002.

[4] M. de Fátima F. L. Rasera, M. V. R. Ballester, A. V. Krusche, C. Salimon, L. A. Montebelo, S. R. Alin, R. L. Victoria, and J. E. Richey, "Estimating the surface area of small rivers in the southwestern Amazon and their role in CO2 outgassing," *Earth Interactions*, vol. 12, pp. 1–16, 2008.

[5] F. M. Hoffman, J. Kumar, R. T. Mills, and W. W. Hargrove, "Representativeness-based sampling network design for the State of Alaska," *Landscape Ecol.*, vol. 28, no. 8, pp. 1567–1586, Oct. 2013.

[6] W. W. Grabowski, "An improved framework for superparameterization," *Journal of the atmospheric sciences*, vol. 61, no. 15, pp. 1940–1952, 2004.

[7] D. Randall, M. Khairoutdinov, A. Arakawa, and W. Grabowski, "Breaking the cloud parameterization deadlock," *Bulletin of the American Meteorological Society*, vol. 84, no. 11, pp. 1547–1564, 2003.

[8] J. Kumar and F. M. Hoffman, "Personal communication," 2013.

[9] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ser. ICML'16. JMLR.org, 2016, p. 478–487.

[10] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.