EdgeAI: How to Use AI to Collect Reliable and Relevant Watershed Data

Authors

Maruti K. Mudunuru[1], Xingyuan Chen[1], Satish Karra[2], Glenn Hammond[1], Peishi Jiang[1], Kurt C. Solander[2], Kalyana B. Nakshatrala[3], Alexander Sun[4], Neeraj Kumar[1], Roelof Versteeg[5], Nikolla P. Qafoku[1], Adam R. Mangel[1], James Stegen[1], Timothy D. Scheibe[1]

[1]Pacific Northwest National Laboratory, Richland, WA 99352, USA.

[2]Los Alamos National Laboratory, Los Alamos, NM 87545, USA.

[3]Department of Civil & Environmental Engineering, University of Houston, Houston, TX 77204, USA.

[4]Bureau of Economic Geology, The University of Texas at Austin, Austin, TX 78713, USA.

[5]Subsurface Insights LLC, Hanover, NH 03755, USA.

Focal Area(s)

Focal areas are on data acquisition and assimilation enabled by AI, advanced methods including experimental/network design/optimization, unsupervised learning (including deep learning), and hardware-related efforts involving AI (e.g., edge computing)

Science Challenge

The transformational science challenge that we address is the following: – *Ensuring, in near real-time, that the data collected from distributed sensor networks is accurate and contains useful information to identify, quantify, and predict watershed and ecosystem dynamic responses to short- and long-term perturbations.*

Rationale

Research needs and challenges: The increased prevalence of extreme disturbance events such as droughts, floods, wildfires, rain-on-snow events, and extreme temperatures is having a profound impact on watershed hydrology and biogeochemical cycles[1] (e.g., timescale in the order of days to years). Additionally, a watershed's hydro-biogeochemistry is also altered considerably by changes in long-term mean climate perturbations such as rising temperatures, changes in the magnitude and frequency of precipitation, earlier snowmelt in mountainous regions, reduced capacity to sequester carbon by the loss of wetlands[2-5], and agricultural intensification (e.g., through enhanced nutrient loading[6]).

To better understand the variable ecosystem response (e.g., biogeochemical stocks and fluxes) under such a wide range of environmental conditions and ecological stressors, a variety of environmental datasets are actively being acquired. These experimental and observational datasets are commonly used in process models in a coupled modeling-experimental (ModEx) approach. The focus is to understand watershed function and key hydro-biogeochemical processes under different environmental and climate stressors. However, there are four major challenges associated with this traditional ModEx approach.

1Q. *The first major challenge is related to the quality of the collected data*. Current data acquisition techniques are frequently performed manually at the site where data is collected. This is often an expensive and time-consuming process, requiring high power output to ensure data collection devices provide a continuous and reliable data stream. Moreover, the acquired data can be of large volume, if sampled at medium- to high-frequency range (e.g., 10s of Hz to kHz). The low sampling densities, gaps in datasets, sensor fouling over time, and signal drifting pose another set of challenges. This reduced data quality from multiple sensors can result in poor sensor netting[7]. That is, predictability of watershed's response is diminished due to poor overlapping coverage from two or more underperforming sensors. Due to these issues, a number of data processing techniques must be applied to fill in data gaps or interpolate data across space and/or time, which leads to high levels of uncertainty. As a result, data validation and data-worth analysis can take several days after acquisition before it is actively integrated into the process models.

2Q. *The second major challenge is related to predictability of a watershed's response (e.g., evolution of microbial activity) under disturbances and long-term perturbations using process-based models*

DOI: https://doi.org/10.2172/1769700
DOE BER, AI4ESP 2021 #AI4ESP1095 (https://www.ai4esp.org)

in near real-time. Existing end-to-end process-based modeling workflows[8-11] to enhance the scientific understanding of hydro-biogeochemical dynamics at different spatial and temporal scales can take several weeks to months due to model complexity and the computational resources required to run such models. As a result, these conventional workflows are not ideal for predicting the immediate consequences and implications of environmental stressors on watershed and ecosystem responses in near real-time.

3Q. *The third challenge is related to the question of when, how, and where to collect data.* It is recognized that watershed behavior is to a large extent driven by hot spots and hot moments. Measuring everywhere all the time is infeasible. Thus, we need a way to decide when and how to measure – and possibly even where.

4Q. *The final challenge is that of how to deal with large data volumes.* For multiple sensing modalities (e.g., multispectral cameras or distributed acoustic sensing) the data volumes which can be collected are substantial. Transmitting this data to a central location for analysis is often infeasible, and so there is interest in data reduction at the sensor in an intelligent manner.

Addressing these key challenges: Our approach to address these four key challenges is to make the sensor nodes to be self-aware and intelligent[12]. This is achieved through our EdgeAI workflow (e.g., edge- and fog-level intelligence)[35-42]. This workflow (Figure-1 and Figure-2) will recognize:

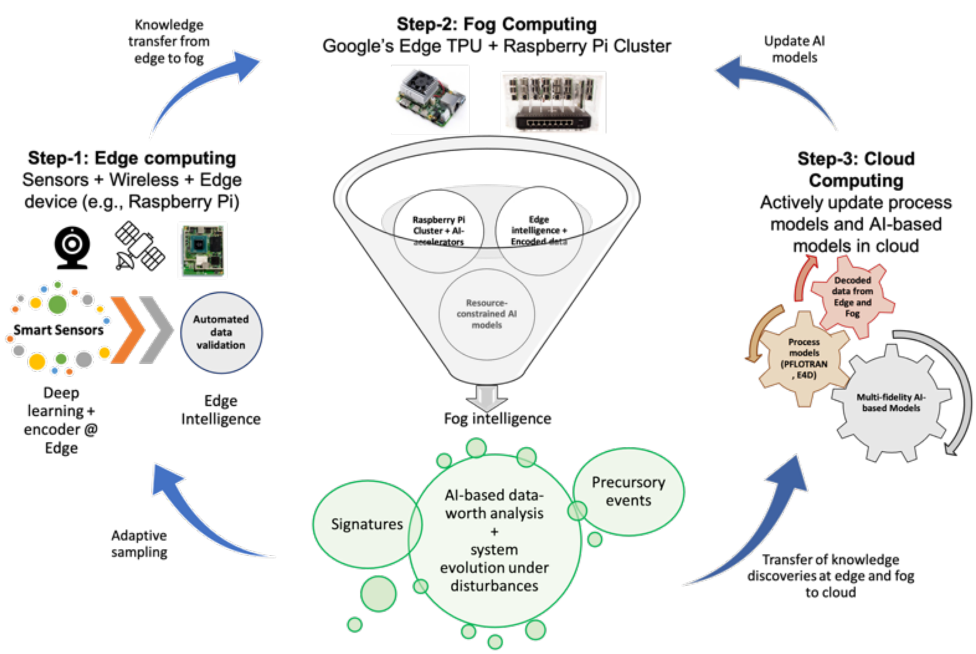

Figure 1 EdgeAI workflow to extract actionable information, discover knowledge, and forecast watershed's response under disturbances

1A. Quality of the collected data – EdgeAI will provide automated streaming analytics in-situ at the point of data collection (e.g., through TensorFlow Extended[31], TensorFlow Lite[30], EdgeML[29], Waggle[32-34]). Automated data quality validation techniques[49] will be used at the sensor nodes and within sensor networks to ensure data is reliable and of good quality. Validation includes data similarity checks and fingerprinting the acquired data. We ensure that distributions in collected data are similar to process representation and model data of extremes. AI-based local data-worth analysis[50,51] will be used to determine if the data contains useful information to detect signatures and underlying patterns. Unsupervised learning and outlier detection methods such as matrix profiling[52], Fingerprint and Similarity Thresholding[53], tensor factorization[54], and autoencoders[55] will be used for pattern recognition. The discovered patterns are then converted to actionable intelligence (e.g., system evolution) at edge- and fog-levels.

2A. Understanding the hydro-biogeochemical response of the watershed under different disturbances – EdgeAI will provide near real-time assimilation of extracted information to improve hydro-biogeochemistry process models (e.g., parameters and microbial functional representations in PFLOTRAN, E4D)[13-19] and forecast the watershed's response under various environmental and climate perturbations.

3A. When, how, and where to collect data – This is achieved by creating a digital twin for watersheds. Through this digital twin, we can optimize the value of controlled disturbance experiments by exploring system behaviors in a digital space. Additionally, measurements at different scales in space could be optimized depending on observed system behaviors, antecedent or forecasted perturbations, and the value of information.

4A. Dealing with large-volumes of data – Through 5G enabled AI@Edge programming models[34] for resource-constrained computing. Multi-fidelity EdgeAI-based models (e.g., Transformer neural networks[20], RNNPool[21]) are trained at edge, fog, and cloud computing devices through self-attention. Recent advances in Raspberry Pi CM4+ and Array of Things (AoT) when combined with Google's Edge TPU allows us to perform ultra-fast inference and vision at edge.

EdgeAI sensor networks interfacing with FAIR data sources: The real-time measurements that we will acquire and assimilate into EdgeAI models include hydrological, biological, geophysical, and geochemical datasets[22,23,24]. Data acquisition uses existing resources such as AmeriFlux Network[25], NGEE-Tropics[26], SPRUCE[27], and WHONDRS[28]. Edge AI can be embedding on these distributed sensor networks through smart computing devices such as Raspberry Pi CM4+. These sensor edge devices also provide a venue to interface with next generation WiFi and 5G[34]. The high-quality data that is acquired from these sensor networks and processed using EdgeAI algorithms will be made reusable and findings reproduceable through FAIR data sources such as ESS-DIVE and WHONDRS.

Why we expect our approach to succeed? The main benefit is that EdgeAI workflow *can transform raw data that is collected at a wide range of frequencies (e.g., ranging from Hz to MHz) on the sensor edge devices into actionable information in a resource-efficient way*[42-48].

Narrative

Scientific and technical description of the opportunities: An EdgeAI workflow[35-42] provides a transformational way for heterogeneous multi-sensor data fusion (e.g., combining geophysical, geochemical and hydrological data sampled at different frequencies) at the edge devices. Moreover, efficiently harnessing the connectivity of intelligent sensors through edge and fog computing will result in an advanced understanding of watersheds under disturbances and extreme events in near real-time (see Figure-2).

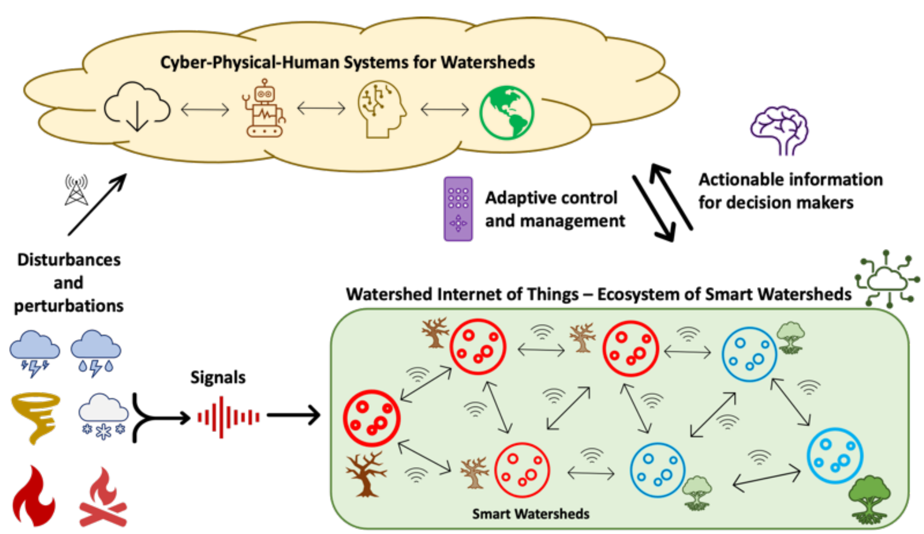

Figure 2 A pictorial description of advanced connectivity of watersheds towards real-time extraction of actionable information, adaptive control, and resource management under disturbances.

Activities that will advance the science: We believe development in data acquisition systems, sensor network design, hardware-related efforts (e.g., AI-accelerators), light-weight AI models (e.g., energy-efficient transformers), and cyber security for edge computing will advanced the proposed science.

EdgeAI: How to Use AI to Collect Reliable and Relevant Watershed Data

Suggested Partners/Experts (Optional):

Laboratory and university partners or experts in the research areas of edge computing, 5G, embedded systems, cyber security, data acquisition, and experts from DOE user facilities may be able to present a related webinar or plenary presentation at a workshop. For instance, these researchers can be Array of Things (AoT) team at Argonne National Laboratory: https://arrayofthings.github.io/

References (Optional)

1. U.S. DOE. 2018. Earth and Environmental Systems Sciences Division Strategic Plan 2018–2023, DOE/SC–0192, U.S. Department of Energy Office of Science (science.osti.gov/-/media/ber/pdf/workshop-reports/2018_CESD_Strategic_Plan.pdf).
2. USGCRP, 2018: Second State of the Carbon Cycle Report (SOCCR2): A Sustained Assessment Report. [Cavallaro, N., G. Shrestha, R. Birdsey, M. A. Mayes, R. G. Najjar, S. C. Reed, P. Romero-Lankao, and Z. Zhu (eds.)]. U.S. Global Change Research Program, Washington, DC, USA, 878 pp., doi: 10.7930/SOCCR2.2018 (https://carbon2018.globalchange.gov/chapter/13/)
3. Carbon Sequestration in Wetlands: http://bwsr.state.mn.us/carbon-sequestration-wetlands
4. A. M. Nahlik, and M. S. Fennessy. 2016. Carbon Storage in U.S. Wetlands. Nature Communications 7:13835.
5. J. Kusler, and J. Christie. 2011. Wetlands and Carbon Storage and Carbon Sequestration. White Paper: Reducing Climate Change Impacts and Promoting Fish and Wildlife: Findings and Recommendations for Biological Carbon Storage and Sequestering. Association of Fish and Wildlife Agencies and the Association of State Wetland Managers.
6. **N. P Qafoku.** 2015. Climate change effects on soils: Accelerated weathering, soil carbon and elemental cycling. Advances in Agronomy: 131, 111-172
7. M. Ilyas, and I. Mahgoub, I. eds., 2004. Handbook of sensor networks: compact wireless and wired sensing systems. CRC press.
8. L. Leonard, and C. J. Duffy. "Automating data-model workflows at a level 12 HUC scale: Watershed modeling in a distributed computing environment." Environmental modelling & software 61 (2014): 174-190.
9. M. V. Muste, D. A. Bennett, S. Secchi, J. L. Schnoor, A. Kusiak, N. J. Arnold, U. Rapolu (2013). End-to-end cyberinfrastructure for decision-making support in watershed management. Journal of Water Resources Planning and Management, 139(5), 565-573.
10. G. Whelan, K. Kim, R. Parmar, G. F. Laniak, K. Wolfe, M. Galvin, M. Molina. "Capturing microbial sources distributed in a mixed-use watershed within an integrated environmental modeling workflow." Environmental modelling & software 99 (2018): 126-146.
11. T. Nyerges, B. Hrishikesh, C. Steinitz, T. Canfield, M. Roderick, J. Ritzman, and W. Thanatemaneerat. "Geodesign dynamics for sustainable urban watershed development." Sustainable Cities and Society 25 (2016): 13-24.
12. N. TaheriNejad, M. A. Shami, and P. D Sai Manoj (2017). Self-aware sensing and attention-based data collection in multi-processor system-on-chips. In 2017 15th IEEE International New Circuits and Systems Conference (NEWCAS) (pp. 81-84). IEEE.
13. PFLOTRAN: https://www.pflotran.org/
14. **G. E. Hammond,** P. C. Lichtner, R. T. Mills (2014). Evaluating the performance of parallel subsurface simulators: An illustrative example with PFLOTRAN. Water resources research, 50(1), 208-228.
15. P. C. Lichtner, G**. E. Hammond**, C. Lu, **S. Karra**, G. Bisht, B. Andre, R. T. Mills, and J. Kumar (2015). PFLOTRAN user manual: A massively parallel reactive flow and transport model for describing surface and subsurface processes (No. LA-UR-15-20403).

16. E4D: https://www.pnnl.gov/projects/e4d
17. T. C. Johnson, **R. J. Versteeg,** A. Ward, F. D. Day-Lewis, A. Revil (2010). Improved hydrogeophysical characterization and monitoring through parallel modeling and inversion of time-domain resistivity andinduced-polarization data. Geophysics, 75(4), WA27-WA41.
18. T. C. Johnson, **G. E. Hammond, X. Chen** (2017). PFLOTRAN-E4D: A parallel open source PFLOTRAN module for simulating time-lapse electrical resistivity data. Computers & Geosciences, 99, 72-80.
19. B. Ahmmed, **M. K. Mudunuru**, **S. Karra**, S. C. James, H. Viswanathan, J. A. Dunbar (2020). PFLOTRAN-SIP: A PFLOTRAN Module for Simulating Spectral-Induced Polarization of Electrical Impedance Data. Energies, 13(24), 6552.
20. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, N. Houlsby (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
21. O. Saha, A. Kusupati, H. V. Simhadri, M. Varma, P. Jain (2020). RNNPool: Efficient Non-linear Pooling for RAM Constrained Inference. arXiv preprint arXiv:2002.11921.
22. S. A. al Hagrey, R. Meissner, U. Werban, W. Rabbel, A. Ismaeil (2004). Hydro-, bio-geophysics. The leading edge, 23(7), 670-674.
23. E. A. Atekwana, L. D. Slater (2009). Biogeophysics: A new frontier in Earth science research. Reviews of Geophysics, 47(4).
24. **A. R. Mangel,** B.A. Lytle, S. M. Moysey, 2015, "Automated high-resolution GPR data collection for monitoring dynamic hydrologic processes in two and three dimensions", The Leading Edge, vol. 2, issue. 34, pp. 190-196, doi: 10.1190/tle34020190.1.
25. AmeriFlux Network: https://ameriflux.lbl.gov/
26. NGEE-Tropics: https://ngee-tropics.lbl.gov/
27. SPRUCE: https://mnspruce.ornl.gov/
28. WHONDRS: https://www.pnnl.gov/projects/WHONDRS
29. EdgeML: https://microsoft.github.io/EdgeML/
30. TensorFlow Lite: https://www.tensorflow.org/lite
31. TensorFlow Extended: https://www.tensorflow.org/tfx
32. Waggle: An open Platform for AI@Edge Computing and Intelligent Sensors https://wa8.gl/science/array-of-things/
33. AoT: Array of Things -- https://arrayofthings.github.io/
34. P. Beckman, C. Catlett, M. Ahmed, M. Alawad, L. Bai, P. Balaprakash, K. Barker, P. Beckman, R. Berry, A. Bhuyan, and G. Brebner (2020). 5G Enabled Energy Innovation: Advanced Wireless Networks for Science, Workshop Report. USDOE Office of Science (SC)(United States).
35. C. J. Talsma, **K. C. Solander, M. K. Mudunuru**, B. Crawford, and M. R. Powell, Frost Prediction using Machine Learning and Deep Neural Network Models for Use on IoT Sensors, under review in IEEE Internet of Things Journal, 2020. LA-UR-20-26558; PNNL-SA-157160.
36. B. Yuan, Y. J. Tan, **M. K. Mudunuru**, O. E. Marcillo, A. A. Delorey, P. M. Roberts, J. D. Webster, C. N. Gammans, **S. Karra**, P. A. Johnson (2019). Using machine learning to discern eruption in noisy environments: A case study using CO2-driven cold-water geyser in Chimayó, New Mexico. Seismological Research Letters, 90(2A), 591-603.
37. **M. K. Mudunuru** (2019). EDGEip - Intelligent Processing at the Edge to Enhance Efficiency, LA-UR-19-29757.
38. **M. K. Mudunuru**, M. B. Cernicek (2018). An INTERNET OF THINGS Commercial Opportunity, LA-UR-18-27304.

39. **M. K. Mudunuru**, V. K. Chillara, **S. Karra**, D. Sinha (2017). Scalable time-series feature engineering framework to understand multiphase flow using acoustic signals. In Proceedings of Meetings on Acoustics 6ICU (Vol. 32, No. 1, p. 055003). Acoustical Society of America.
40. **M. K. Mudunuru** (2019). IoGES: Internet of Things for Geophysical and Environmental Sensing; LA-UR-19-30467; 2019. DOI: 10.13140/RG.2.2.32470.40006
41. **M. K. Mudunuru** (2017). IoGET: Internet of Geophysical and Environmental Things (No. LA-UR-17-25560). Los Alamos National Lab.(LANL), Los Alamos, NM (United States).
42. J. R. Tempelman, **M. K. Mudunuru, S. Karra**, A. J. Wachtor, B. Ahmmed, E. B. Flynn, J.B. Forien, G. Guss, N. Calta, J. P. DePond, and M. Matthews. "Machine Learning on Acoustic Measurements to Predict Pore Formation in Additive Manufacturing." submitted to Additive Manufacturing, LA-UR-20-26174; PNNL-SA-157251
43. N. V. Jagtap, **M. K. Mudunuru**, and **K. B. Nakshatrala**. A deep learning modeling framework to capture mixing patterns in reactive-transport systems. arXiv preprint arXiv:2101.04227 (2021).
44. **K. B. Nakshatrala** and M. S. Joshaghani (2019. On interface conditions for flows in coupled free-porous media. Transport in Porous Media, 130(2), pp.577-609.
45. **A. Y. Sun,** Z. Zhong, H. Jeong, and Q. Yang, (2019). Building complex event processing capability for intelligent environmental monitoring. Environmental Modelling & Software, 116, pp.1-6.
46. **A. Y. Sun,** and B. R. Scanlon (2019). How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. Environmental Research Letters, 14(7), 073001.
47. Z. Kakalia, C. Varadharajan, E. Alper, E. Brodie, M. Burrus, R. Carroll, D. Christianson, V. Hendrix, M. Henderson, S. Hubbard, D. Johnson, **R. Versteeg**, K. Williams, and D. Agarwal, The East River Community Observatory Data Collection: Diverse, multiscale data from a mountainous watershed in the East River, Colorado, Authorea. October 01, 2020.
48. C. Varadharajan, D. A. Agarwal, W. Brown, M. Burrus, R. W. H. Carroll, D. S. Christianson, B. Dafflon, D. Dwivedi, B. J. Enquist, B. Faybishenko, A. Henderson, M. Henderson, V. C. Hendrix, S. S. Hubbard, Z. Kakalia, A. Newman, B. Potter, H. Steltzer, **R. Versteeg**, K. H. Williams, C. Wilmer, and Y. Wu, (2019), Challenges in Building an End-to-End System for Acquisition, Management, and Integration of Diverse Data From Sensor Networks in Watersheds: Lessons From a Mountainous Community Observatory in East River, Colorado. IEEE Access, 7, pp.182796-182813.
49. N. Polyzotis, M. Zinkevich, S. Roy, E. Breck, and S. Whang (2019). Data validation for machine learning. Proceedings of Machine Learning and Systems, 1, pp.334-347.
50. S. Finsterle, (2015). Practical notes on local data-worth analysis. WRR, 51(12), 9904-9924.
51. Y. Li and H. S. Abdel-Khalik (2021) Data trustworthiness signatures for nuclear reactor dynamics simulation. Progress in Nuclear Energy, 133, p.103612.
52. C.C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, D. F. Silva, A. Mueen, and E. Keogh (2016) Matrix profile I: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets. In 2016 IEEE (pp. 1317-1322).
53. C.E., Yoon, O. O'Reilly, K. J. Bergen, and G. C. Beroza (2015). Earthquake detection through computationally efficient similarity search. Science advances, 1(11), p.e1501057.
54. V. V. Vesselinov, **M. K. Mudunuru**, **S. Karra**, D. O'Malley, and B. S. Alexandrov (2019). Unsupervised machine learning based on non-negative tensor factorization for analyzing reactive-mixing. Journal of Computational Physics, 395, pp.85-104.
55. P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, P.A., 2008, July. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning (pp. 1096-1103).