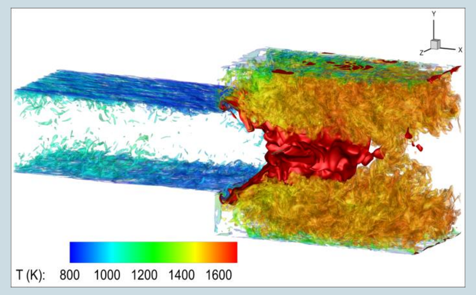
Sandia National Laboratories

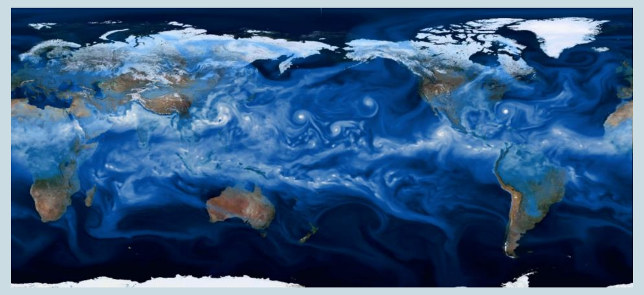
In-Situ Machine Learning for Intelligent Data Capture

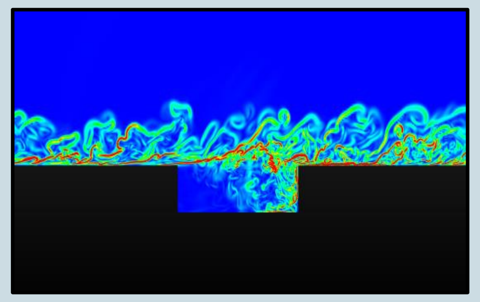
Warren L. Davis IV(wldavis@sandia.gov), Hemanth Kolla, Irina Tezaur, Max Carlson

Motivation

- DOE research often involves discovering new, “interesting” events in high-fidelity physics-based HPC simulations.



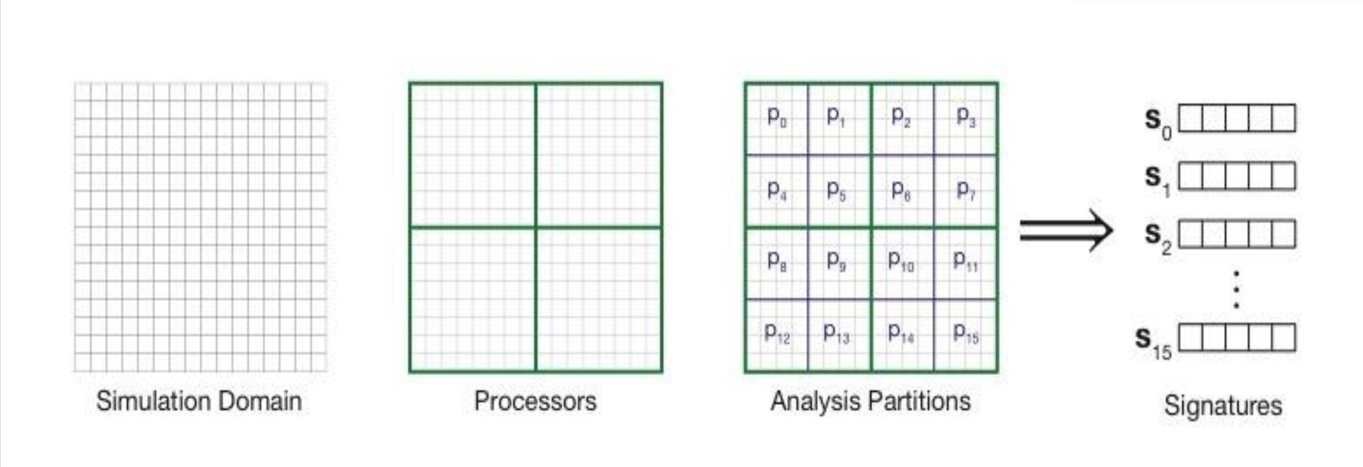




- Standard anomaly detection algorithms are limited, requiring the capture of all the data for post processing, or requiring high-bandwidth in-situ communication.
- Our goal is to create a more efficient way to detect the anomalies *in-situ,* thus facilitating more precisely targeted event capture.

Signatures, Measures, and Decisions

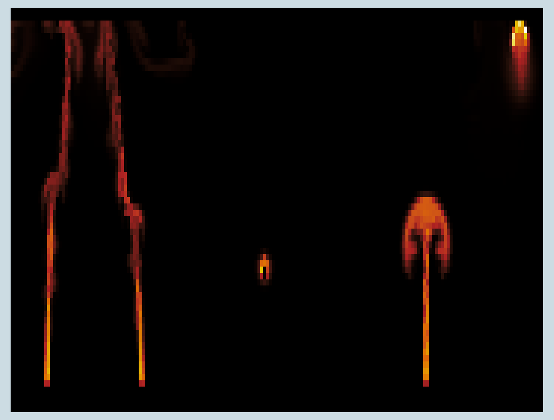
- We divide the simulation domain into analysis partitions, then create compressed representations of simulation data, called **signatures.**



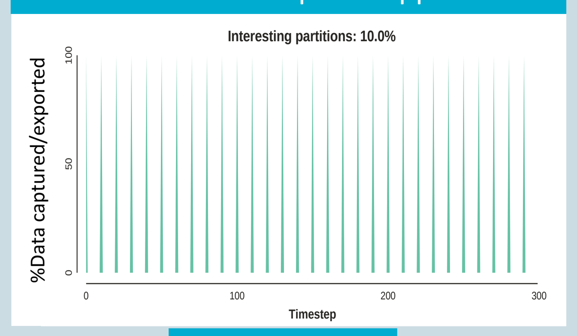
- We then utilize **measures**, which are representations of how close signatures are to each other. Measures can be off-the-shelf algorithms, such as DBSCAN, or specially crafted to work with specific signatures.
- Finally, we use **decisions**, which are a mapping from measures to partitions that may be ”interesting”

Results Summary

We injected randomly distributed anomalies into a buoyant fluids simulation as an initial test of the efficiency of our approach versus the conventional approach of sparse snapshotting. Our methods **captured far more of the anomalies with only slightly higher data export.** Similar results were found in other simulations, such as combustion and cyclone detection.

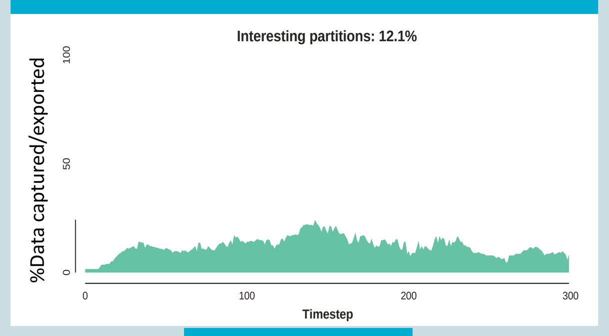


Conventional Snapshot Approach



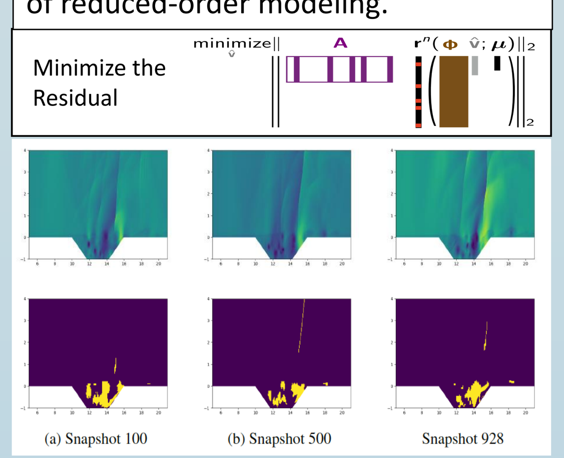
Recall: 10%

In-Situ Detection

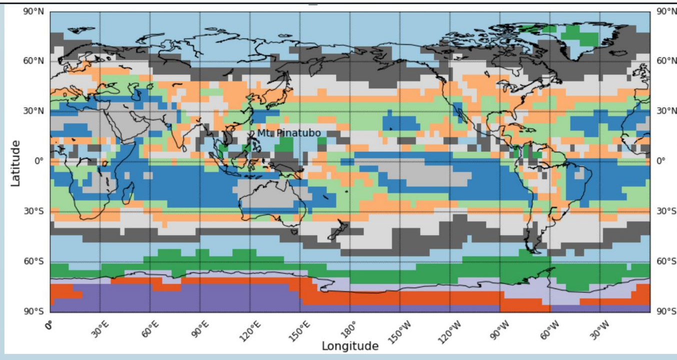


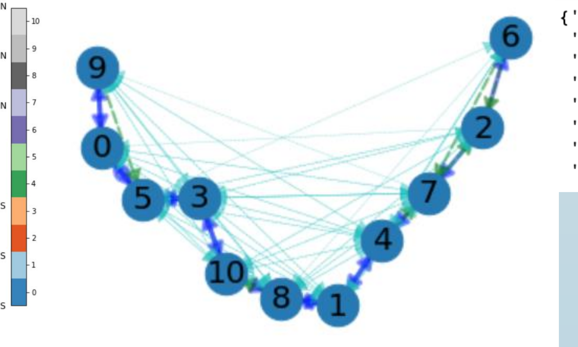
Recall: 98%

We are exploring the use of ISML algorithms in the hyper-reduction step of reduced-order modeling.



We are currently using ISML algorithms in a different fashion, altogether, in the Grand Challenge project “CLimate impact: Determining Etiology thRough pAthways” (CLDERA, PI: Diana Bull). Rather than finding things like anomalies, we are using the algorithms to cluster regions of similar statistical behavior. We can then trace the evolution of cluster partitions through time, tracing chains of events and probabilities of transitions. Shown below is the clustering for a simulation of the eruption of Mt. Pinatubo. To the right of the snapshot is the cluster evolution graph and a sample of the transition probabilities.





{'1, 8, 1, 4, 1': [534, 0.19553277187843282],
'1, 8, 1, 4, 2': [1472, 0.5389967045038447],
'1, 8, 1, 4, 3': [3, 0.0010984987184181618],
'1, 8, 1, 4, 5': [4, 0.0014646649578908824],
'1, 8, 1, 4, 6': [14, 0.005126327352618088],
'1, 8, 1, 4, 7': [660, 0.2416697180519956],
'1, 8, 1, 4, 8': [41, 0.015012815818381545],
'1, 8, 1, 4, 10': [3, 0.0010984987184181618]}

Impacts & Successes to date

- Publications

  - T. M. Shead, I. K. Tezaur, W. L. Davis IV, M. L. Carlson, D. M. Dunlavy, E. J. Parish, P. J. Blonigan, J. Tencer, F. Rizzi, and H. Kolla, “A novel in situ machine learning framework for intelligent data capture and event detection,” in in Lecture Notes in Energy, vol 44. Springer, Cham. January, 2023.
  - A. Konduri, H. Kolla, W. P. Kegelmeyer, T. M. Shead, J. Ling, and W. L. Davis IV, “Anomaly detection in scientific data using joint statistical moments,” J. Comput. Phys., vol. 387, pp. 522–538, 2019.
  - Julia Ling, W. Kegelmeyer, A. Konduri, H. Kolla, K. Reed, T. M. Shead, W. L. Davis IV. (2017). Using feature importance metrics to detect events of interest in scientific computing applications. 55-63.

- Presentations

  - W. L. Davis IV, H. Kolla, I. Tezaur, and M. Carlson, “In-situ machine learning (ISML) for climate simulation analysis and discovery,” in European Seminar on Computing (ESCO 2022), Pilsen, CZ, 2022.
  - W. L. Davis IV, “In-situ machine learning for intelligent data capture on exascale platforms,” Presentation, Platform for Advanced Scientific Computing (PASC) Conference, Geneva, Switzerland, 2021.
  - W. L. Davis IV, “The potential of integrated machine learning algorithms for tropical cyclone detection in advanced climate modeling,” Poster Presentation, American Geophysical Union (AGU) Fall Conference, Washington, D.C., December, 2019.

- Other
  - Open Source Software: https://github.com/sandialabs/isml
  - Integral part of the CLDERA Grand Challenge LDRD, which seeks to discover source-to-impact pathways in climate simulations

ISML was funded by DOE/ASCR originally for ~$1.8 M for 4 years. After this, it was renewed for Phase 2 (year 3 of 4) at ~$2.4 M.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology &Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-NA0003525.
SAND No. _____________

Sandia National Laboratories