This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2020-7823C

In-Situ Machine Learning for Intelligent Data Capture for Exascale Platforms

Machine Learning and Deep Learning Conference 2020

Presented By: Gabriel Popoola

Collaborators: Warren L. Davis IV, Tim Shead, Hemanth Kolla, Kevin Reed, Philip Kegelmeyer

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.

Overview

- Problem Summary
- Framework and Approach
- Data and Applications
- Results
- Conclusions

Problem

Problem Summary

- Scientific computation often involves running computationally intense simulation on HPC
- With the goal being the identification and detection of events of interest, the current strategies for detection are lacking due to the I/O overhead that prevents all data from being written out
- Given that the current HPC Simulation strategy for event and anomaly detection involves saving data at regular intervals, two primary problems surface as a result:
    1. Writing at infrequent intervals leads to missed events and/or loss of critical information
    2. When information is lost, it can only be regained by re-running the simulation and adjusting the save intervals

Framework and Approach

Signatures, Measures, Decisions

Framework – Signatures, Measures, Decisions

- **Project:** ASCR In-Situ Machine Learning for Intelligent Data Capture on Exascale Platforms (ISML)
- The *Signature-Measures-Decision* framework is a generalized way of performing unsupervised anomaly detection
- Data space is defined as the dimensional space that contains all points in the dataset
- Partitions are defined as sections (typically equal sized) of the data space that may or may not contain data points
- The goal of the approach is to detect events of interest—any local (within a partition) occurrences that differ significantly from the occurrences in other partitions or in the same partition at different time steps

Approach – Signatures, Measures, Decisions

Signatures are a compressed way of representing the data in a partition; the representation should contain the crucial aspects of the data such that spatial and/or temporal changes can be detected

Examples:

- Mean
- Min-max
- SVD[9]

Approach – Measures

Measures are functions applied to signatures to detect changes across space or time; spatial measures compare signatures of different partitions at one time and temporal measures compare signatures of a single partition over time

Examples:

- Maximum change (temporal)
- SVD_9 (temporal)
- Mean squared distance (spatial)
- Signature scaling (spatial)
- DBSCAN

Approach – Decisions

Decisions are functions that determine whether the measure should be flagged as anomalous

Examples:

- Threshold
- Compound threshold
- Memory
- Percentile

Data and Applications

Spatial and Temporal

Data: Spatial and Temporal Event/Activity Detection

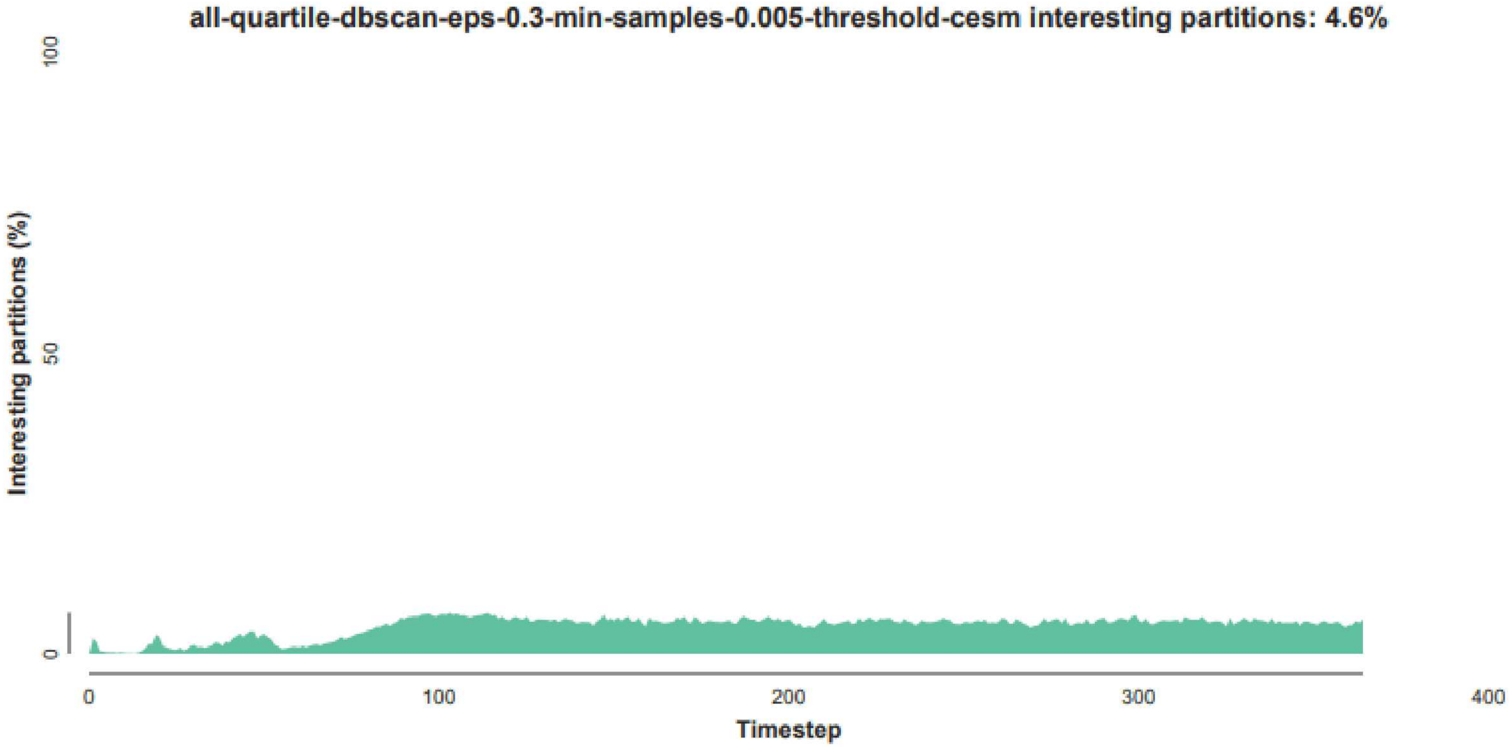
Simulation Case	Description	Features
Community Earth System Model (CESM)	Climate model simulation with events of interest being a set of cyclones throughout the simulation space	• Total Precipitation Rate • Surface Pressure • Lowest model level water vapor mixing ratio • Atmospheric air temperature • Lowest model level zonal wind • Lowest model level meridional wind • Atmospheric reference height
Mantaflow	Fluid model simulation with events of interest being the activity/changes throughout the simulation space	• Velocity X • Velocity Y • Density • Pressure

Results

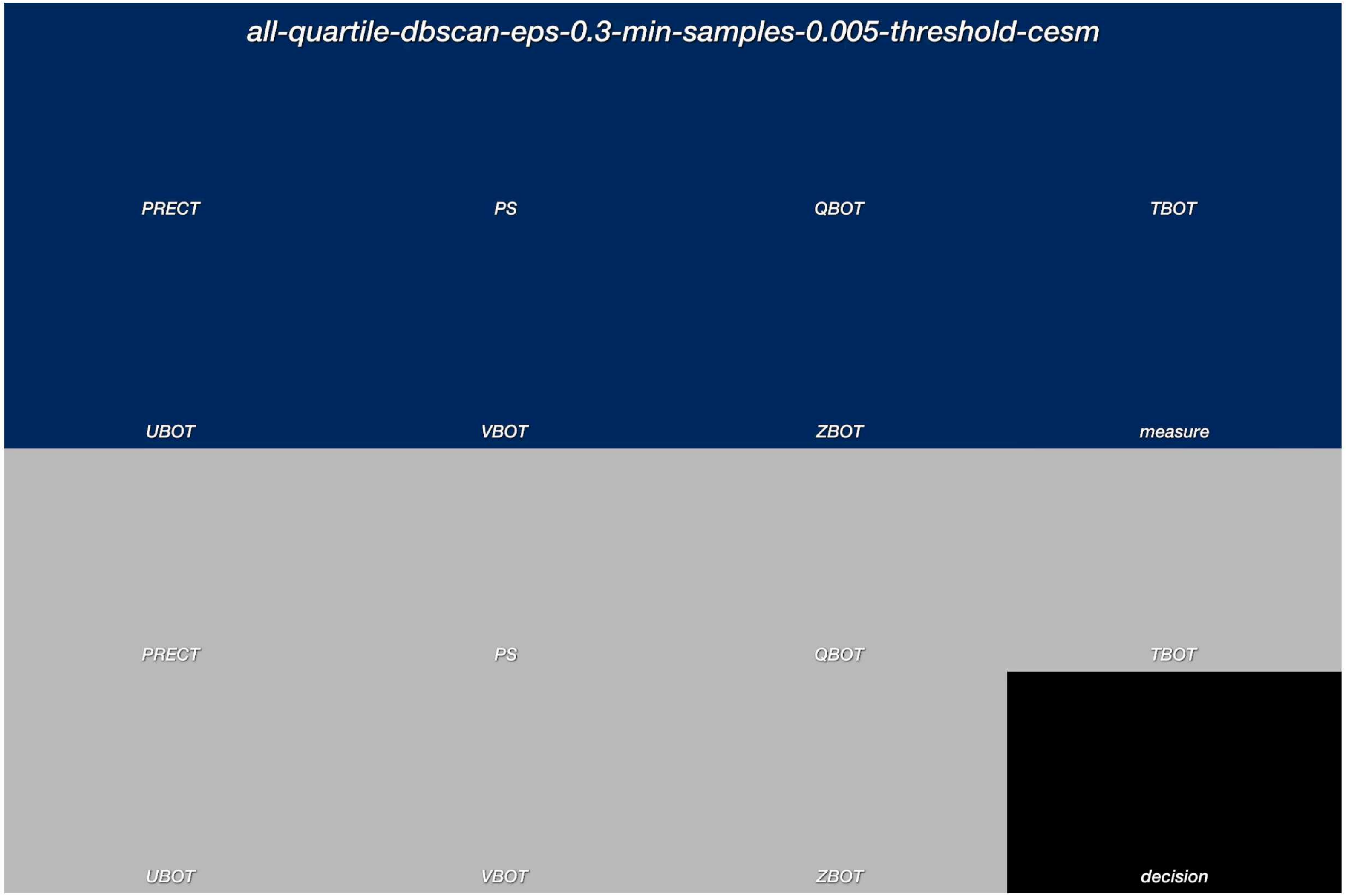
Spatial and Temporal

Results

Percentage of Data Saved Over Time

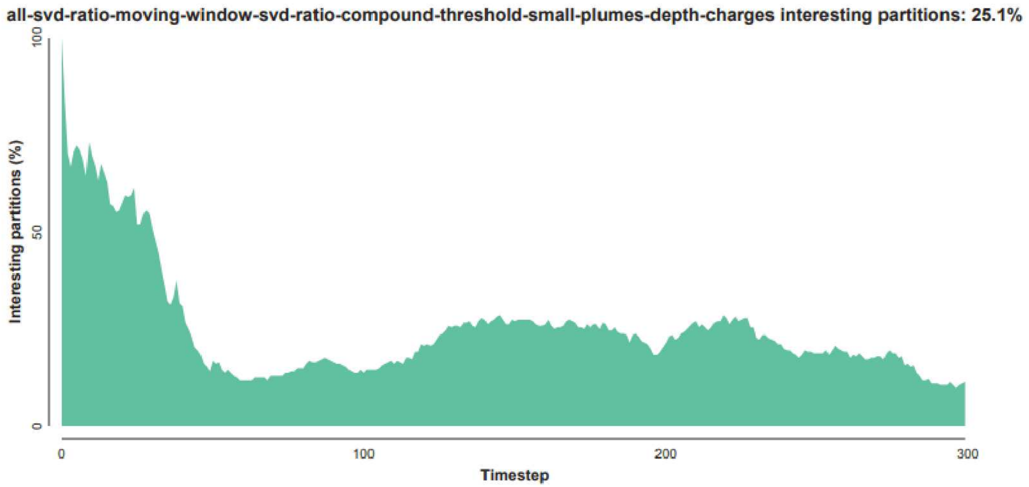


Results: Spatial Event/Activity Detection – Cyclones (CESM)

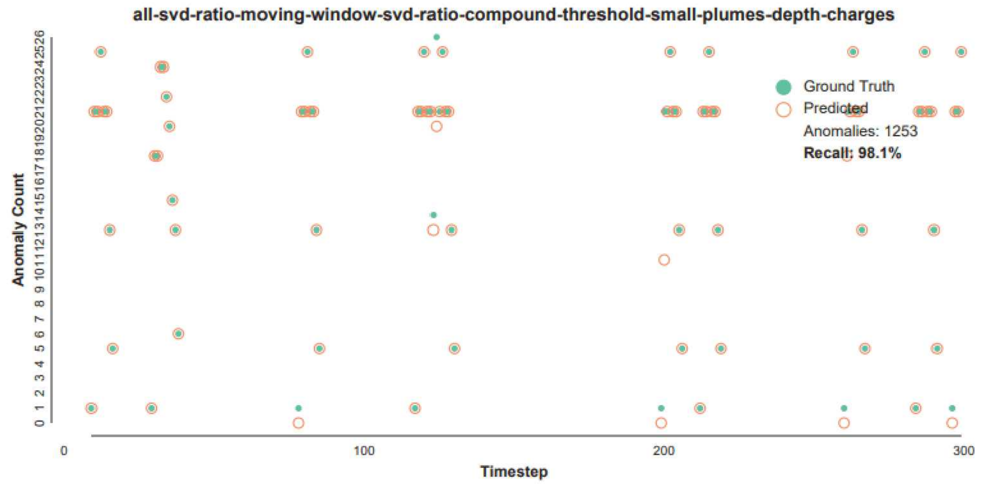


Results

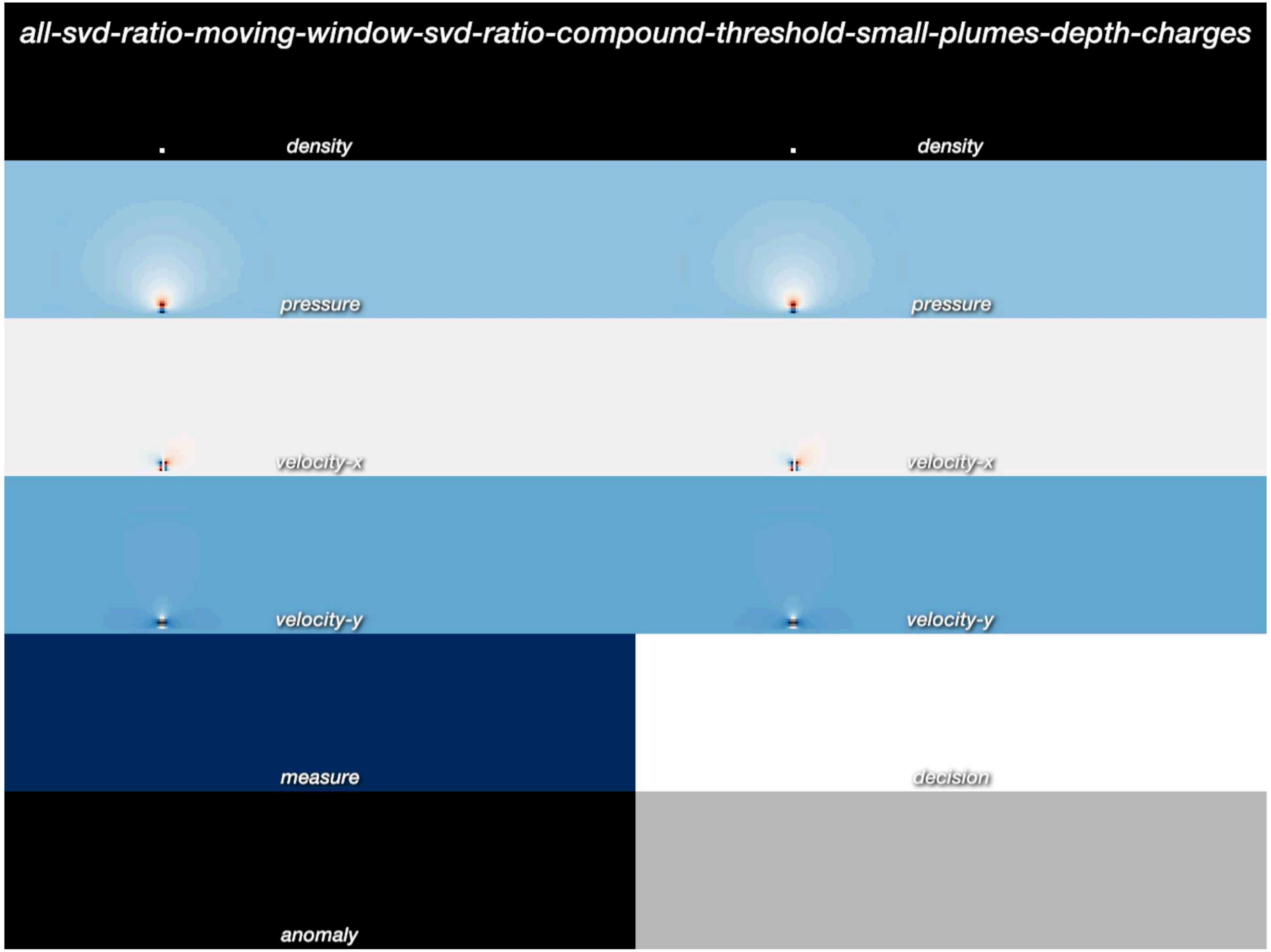
Percentage of Data Saved Over Time



Recall



Results: Temporal Event/Activity Detection – Smoke Plumes



Conclusion

- The framework is capable of detecting events of interest with high and similar accuracy, whether detecting events spatially or temporally
- Though the experiments have been run on simulation data so far, the approach can generalized to other fields and application domains (cyber security, satellite image analysis)
- A major key in applying the framework to other areas is the ability to extract quantifiable features from the data
- The in-situ detection is efficient and accurate

Questions?

Relevant Literature

1. Julia Ling, W. Philip Kegelmeyer, Konduri Aditya, Hemanth Kolla, Kevin A. Reed, Timothy M. Shead, and Warren L. Davis. Using feature importance metrics to detect events of interest in scientific computing applications.2017 IEEE 7th Symposium on Large Data Analysis and Visualization (LDAV), 2017.

2. Subutai Ahmad, Alexander Lavin, Scott Purdy, and Zuha Agha. Unsupervised real-time anomaly detection for streaming data. Neurocomputing, 262:134 – 147, 2017. Online Real-Time Learning Strategies for Data Streams.

3. Herman Aguinis, Ryan K. Gottfredson, and Harry Joo. Best-practice recommendations for defining, identifying, and handling outliers. Organizational Research Methods,16(2):270–301, 2013.

4. Anomaly Detection, a short tutorial using Python - http://aqibsaeed.github.io/2016-07-17-anomaly-detection/

5. Wiliem, Arnold & Madasu, Vamsi & Boles, Wageeh & Yarlagadda, Prasad. (2012). A Context Space Model for Detecting Anomalous Behaviour in Video Surveillance. Proceedings of the 9th International Conference on Information Technology, ITNG 2012. 18-24. 10.1109/ITNG.2012.11.

6. Anomaly Detection Presented by: Anupam Das CS 568MCC Spring 2013

7. Davis IV, Warren Leon; Shead, Timothy M.; Kolla, Hemanth; Reed, Kevin; Kegelmeyer, Philip; Popoola, Gabriel. "In-Situ Machine Learning for Intelligent Data Capture on Exascale Platforms." Artificial Intelligence for Robust Engineering & Science Workshop (AIRES), January 2020

8. Rasmus Bro, Evrim Acar, and Tamara G. Kolda. Resolving the sign ambiguity in the singular rvalue decomposition. Journal of Chemometrics, 22(2):135–140, February 2008.

9. S. Koelstra, C. Muhl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt,and I. Patras. Deap: A database for emotion analysis using physiological signals. IEEETransactions on Affective Computing, 3(1):18–31, Jan 2012.

10. James Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. Journal of Research in Personality, 11:273–294, 09 1977.