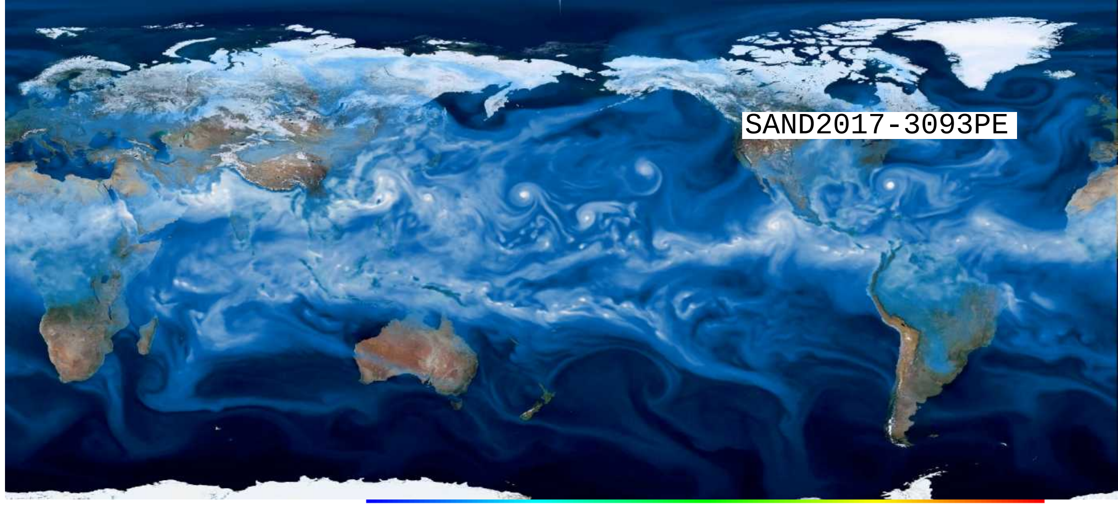
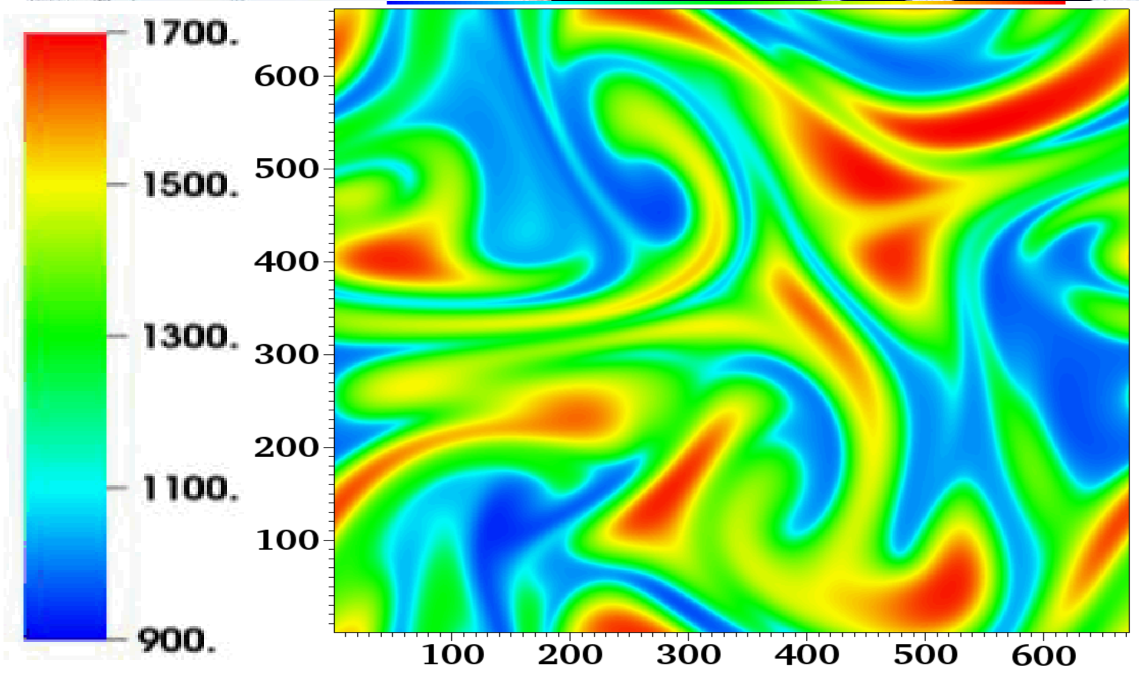
In-Situ Machine Learning for Intelligent Data Capture on Exascale Platforms





Warren L. Davis IV

March 27, 2017

U.S. DEPARTMENT OF ENERGY | Office of Science

Sandia National Laboratories is a multi-mission laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000

Sandia National Laboratories

Stony Brook University

Team

- Warren L. Davis IV (PI-Sandia)
- Kevin Reed (PI-Stony Brook)
- Danny Dunlavy
- Philip Kegelmeyer
- Hemanth Kolla
- Aditya Konduri

- Julia Ling (Former)

Problem

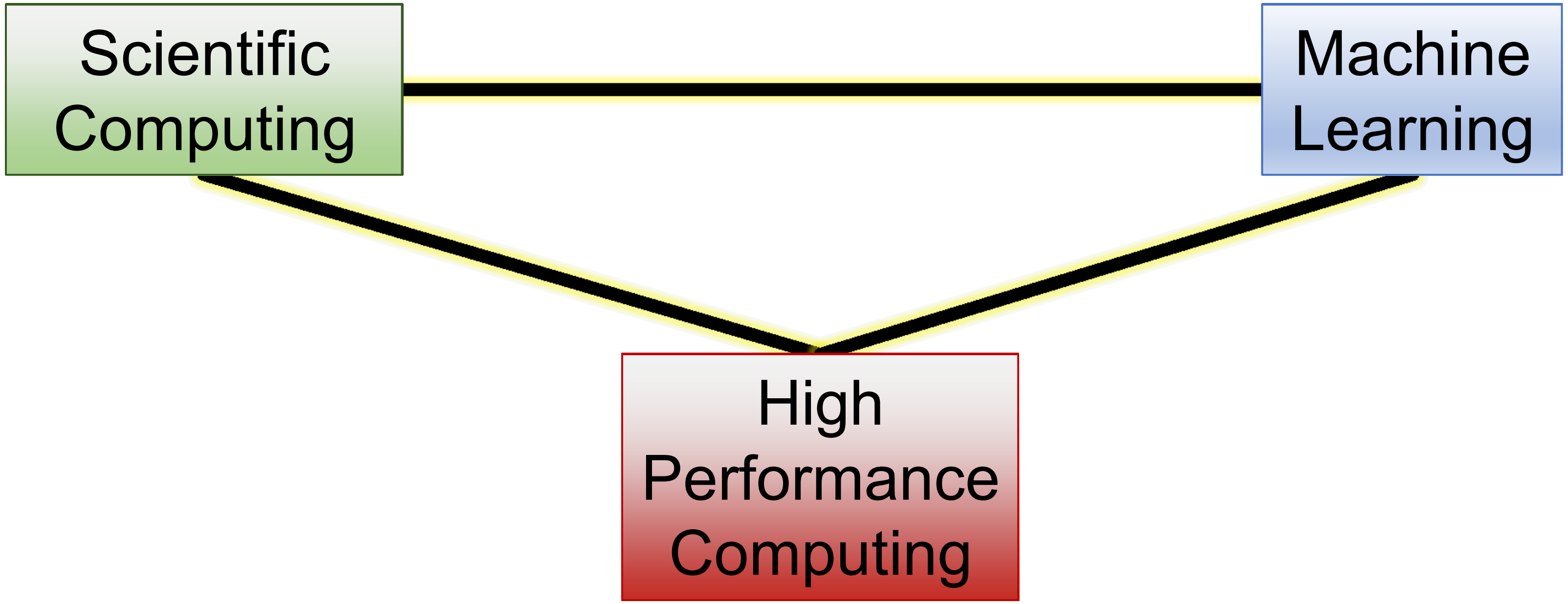
- Scientific computation often involves running computationally intense simulations on HPC
- Goal is to find interesting events (e.g., auto-ignition, cyclones)
- Critical events Current HPC Simulation strategy for detection of events and anomalies involves saving data to disk at regular intervals.

- Overhead for I/O is large
  - Writing everything is too expensive
  - Writing at infrequent intervals may lead to missed events, or loss of ciritical information
  - Lost information can only be regained by rerunning the simulations and adjusting the save interval.

Research Goals

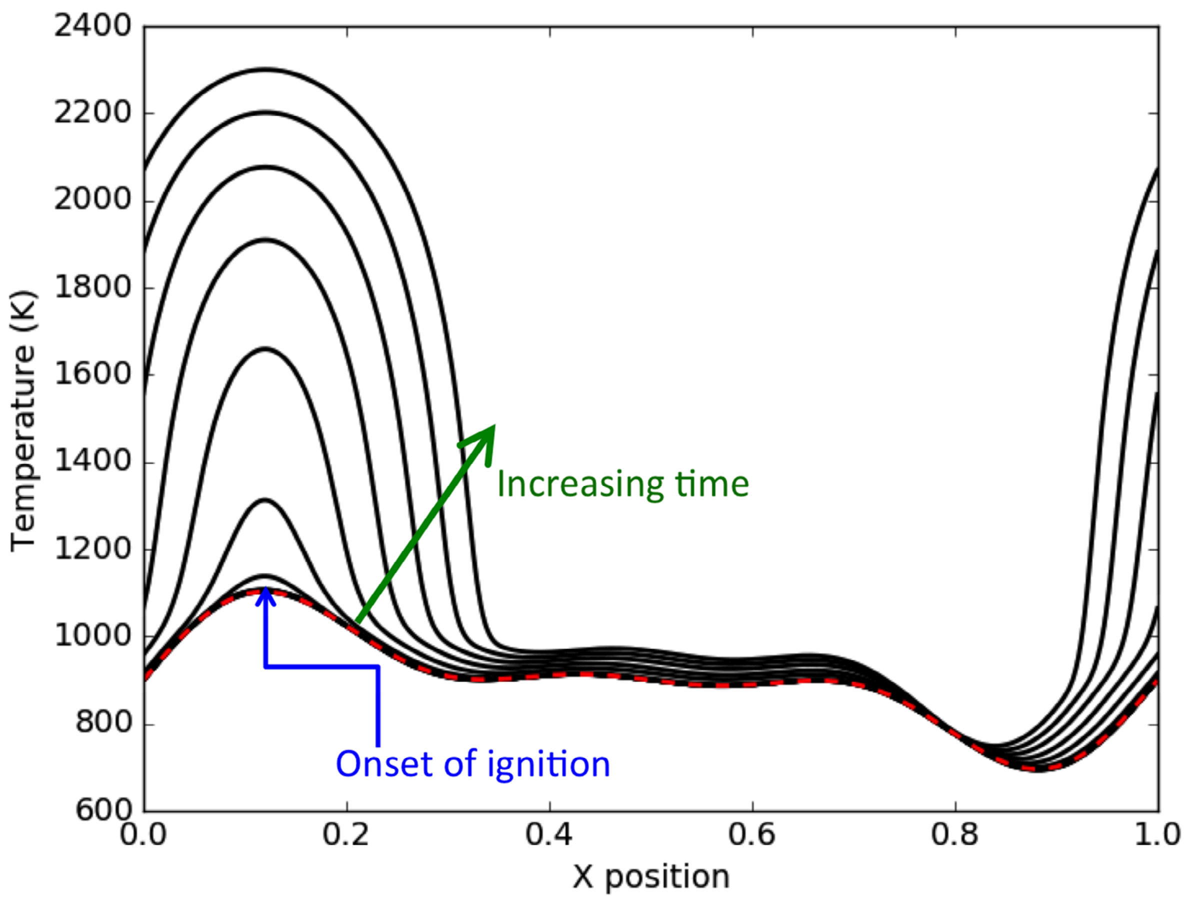
- Develop efficient distributed machine learning and anomaly detection algorithms to enable intelligent data capture.
- These algorithms will be used to determine localized events of interest *in situ*, and the data will be selectively saved at the relevant time steps and spatial locations.
- The machine learning techniques will be implemented and validated on two test cases: auto-ignition in a combustion simulation and extreme weather prediction in a climate simulation.

Primary Components



Auto-Ignition

- Modeled using S3D
- 17 state variables
  - 12 species concentrations
  - 3 velocity components
  - Temperature
  - Pressure
- Temperature profile prescribed as a sum of sines



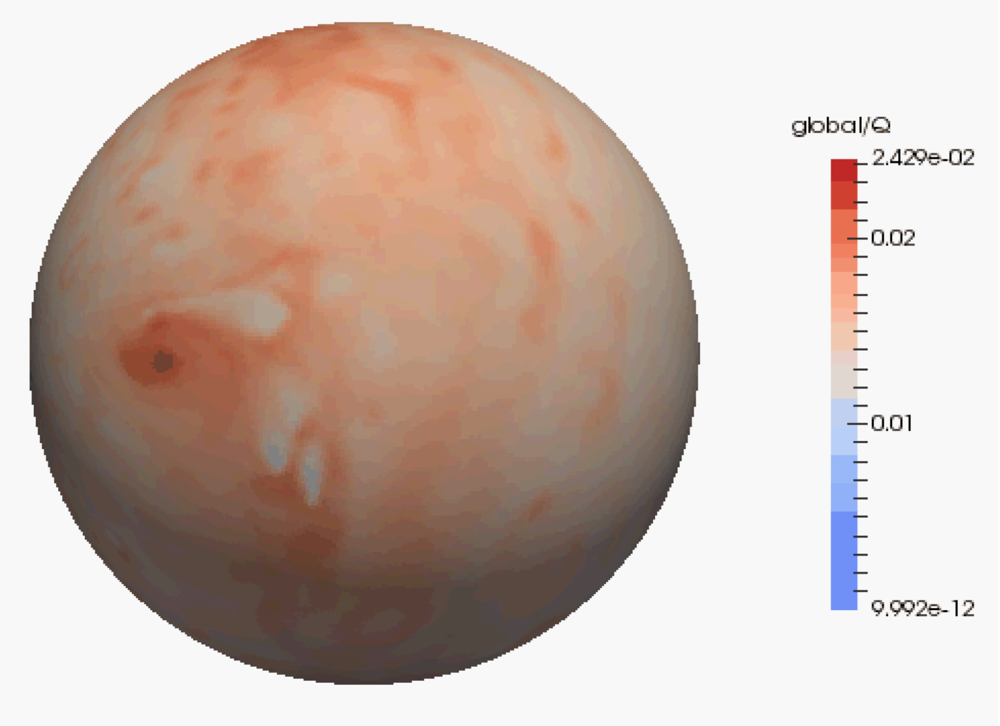
U.S. DEPARTMENT OF ENERGY | Office of Science

Sandia National Laboratories

Stony Brook University

Climate Modeling

- Test-case : Idealized tropical cyclones
- National Center for Atmospheric Research's (NCAR) and Department of Energy (DOE) supported Community Atmosphere Model version (CAM 5).
  - Horizontal resolutions of ~100 km and ~25 km
  - Atmosphere only

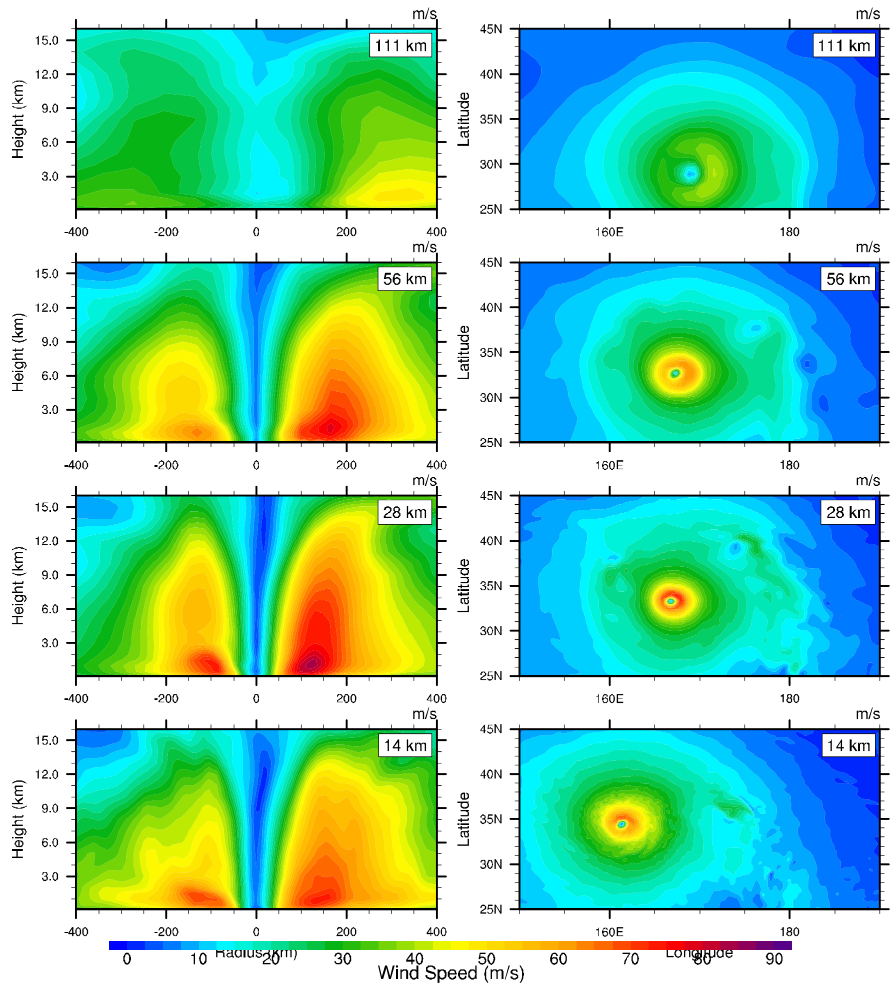


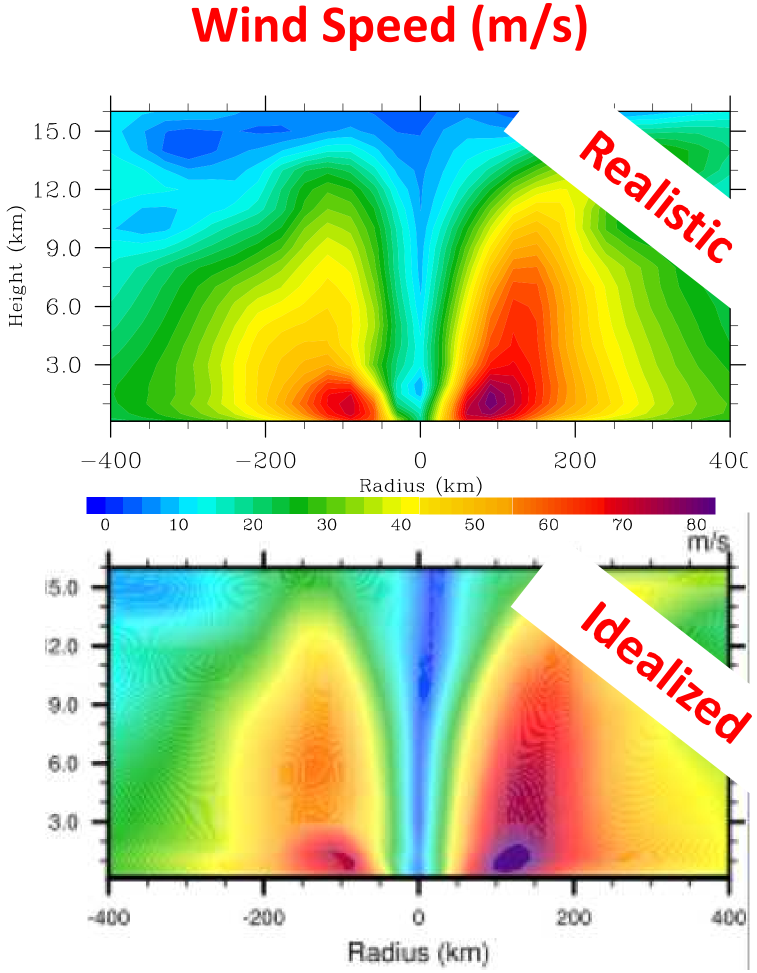
U.S. DEPARTMENT OF ENERGY | Office of Science

Sandia National Laboratories

Stony Brook University

Climate Modeling (cont.)





Idealized model captures most of the interesting aspects that we are trying to detect with ML

Machine Learning

- Anomaly/Change-point Detection

- Desired algorithm attributes
    - Generalizability
    - Unsupervised
    - Low communication overhead
    - Online capability for streaming data

Machine Learning (cont.)

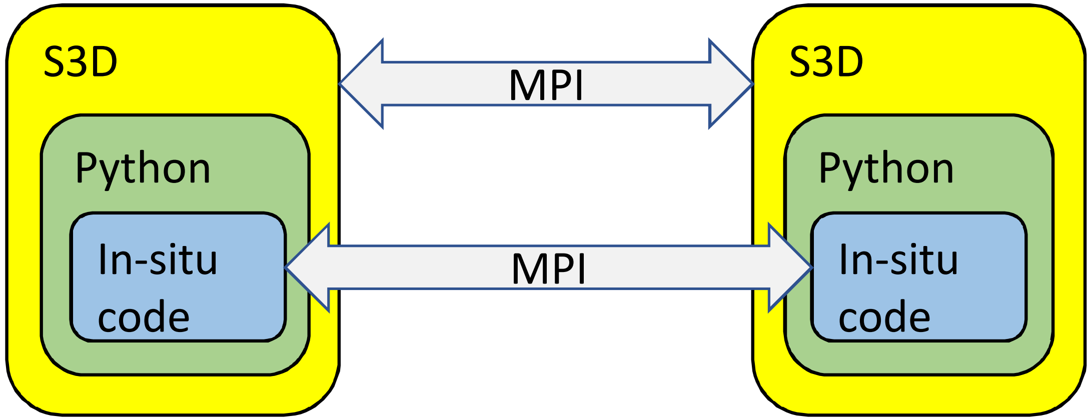
- Built a suite of pre-existing and newly implemented algorithms suitable for integration/experimentation

SVM	PCA	Density Estimation Trees
K-Means	Kernel Density Estimation	Local Outlier Factors
Various distances	Velocity Density Estimation	

Isolation Forests
Isolation Nearest Neighbor Ensembles
Random Subspace Forests
Density Estimation Forest

High Performance Computing

- S3D
  - Scalable parallel direct numerical simulation reacting flow solver used throughout Sandia and the DoE
- Developers created new in-situ capability in S3D
  - Embedded Python interpreter
  - Allows us to execute interpreted code in-situ with full MPI capability



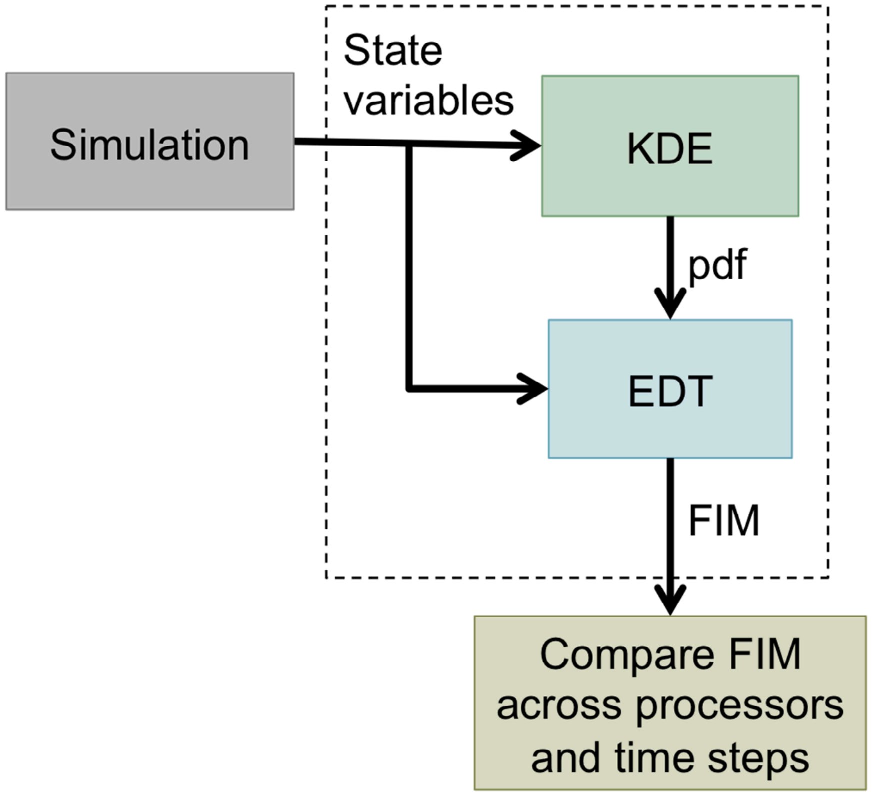
- Allows tight integration with the combustion team
  - Early and deep integration eliminated the need for Mantevo Mini-Apps

Experiments

- Preliminary experiments on auto-ignition and climate models
  - Began in parallel with HPC interface development
  - Using pre-generated data, down-sampled in time
- Moderately successful with existing algorithms
  - Density estimation-related techniques are not as robust as needed
  - Features spanning multiple mesh points and feature drift aren't handled well
- Modified ensemble methods to reduce communication (sparse/performance-based updating)

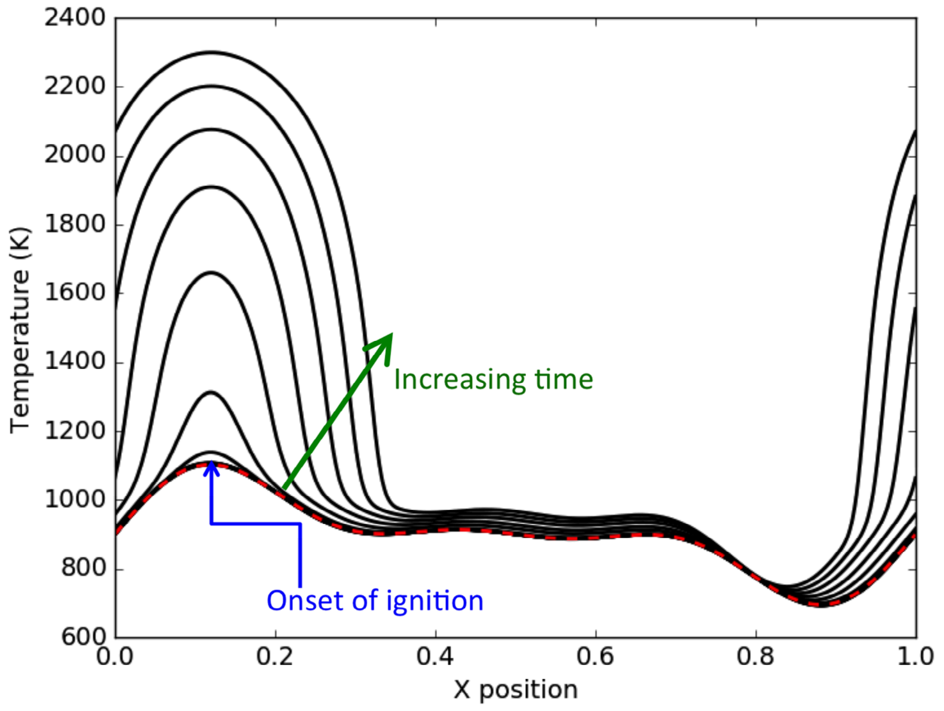
Feature Importance Event Detection Algorithm (FIEDA)

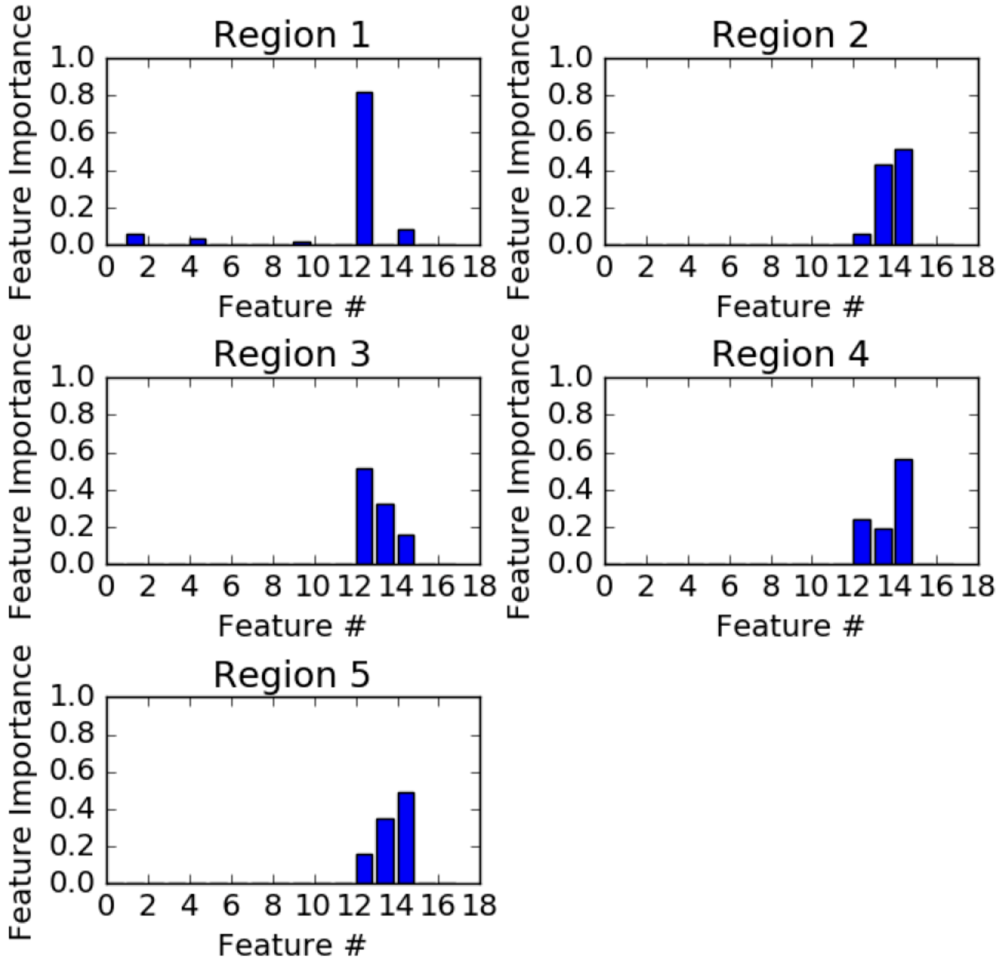
- Use Kernel Density Estimation (KDE) to determine a probability density function (PDF) over the state variables on a processor
- Use Ensemble of Decision Trees (EDT) Regressor to predict the PDF given the state variables
- Extract feature importance metrics (FIM) from the ensemble
- Compare the FIM
  - Across processors (spatial, *M1*)
  - Across time steps (temporal, *M2*)



Auto-Ignition Results

FIM for regions on the onset of ignition





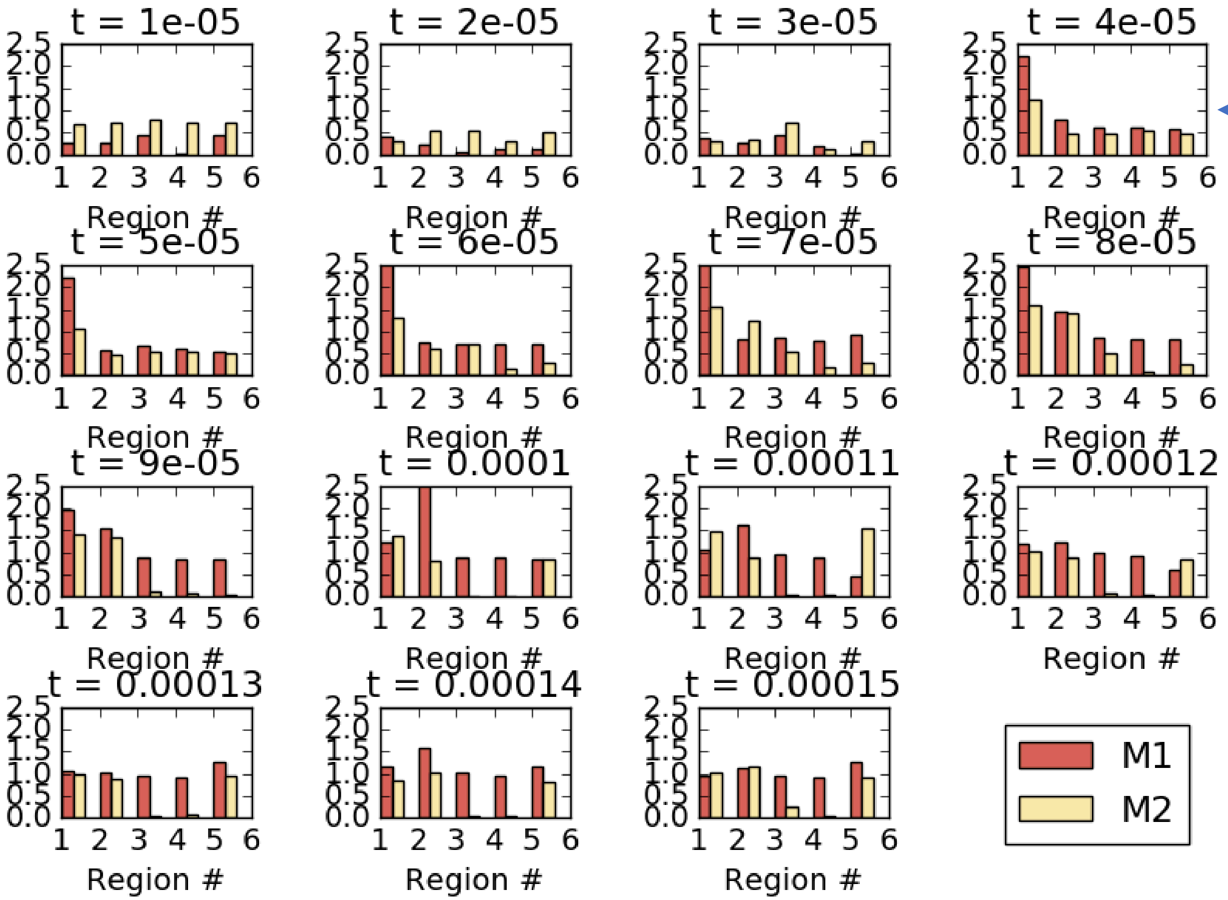
U.S. DEPARTMENT OF ENERGY | Office of Science

Sandia National Laboratories

Stony Brook University

Auto-Ignition Results (cont.)

M1 and M2 values across the 5 regions and 12 time steps

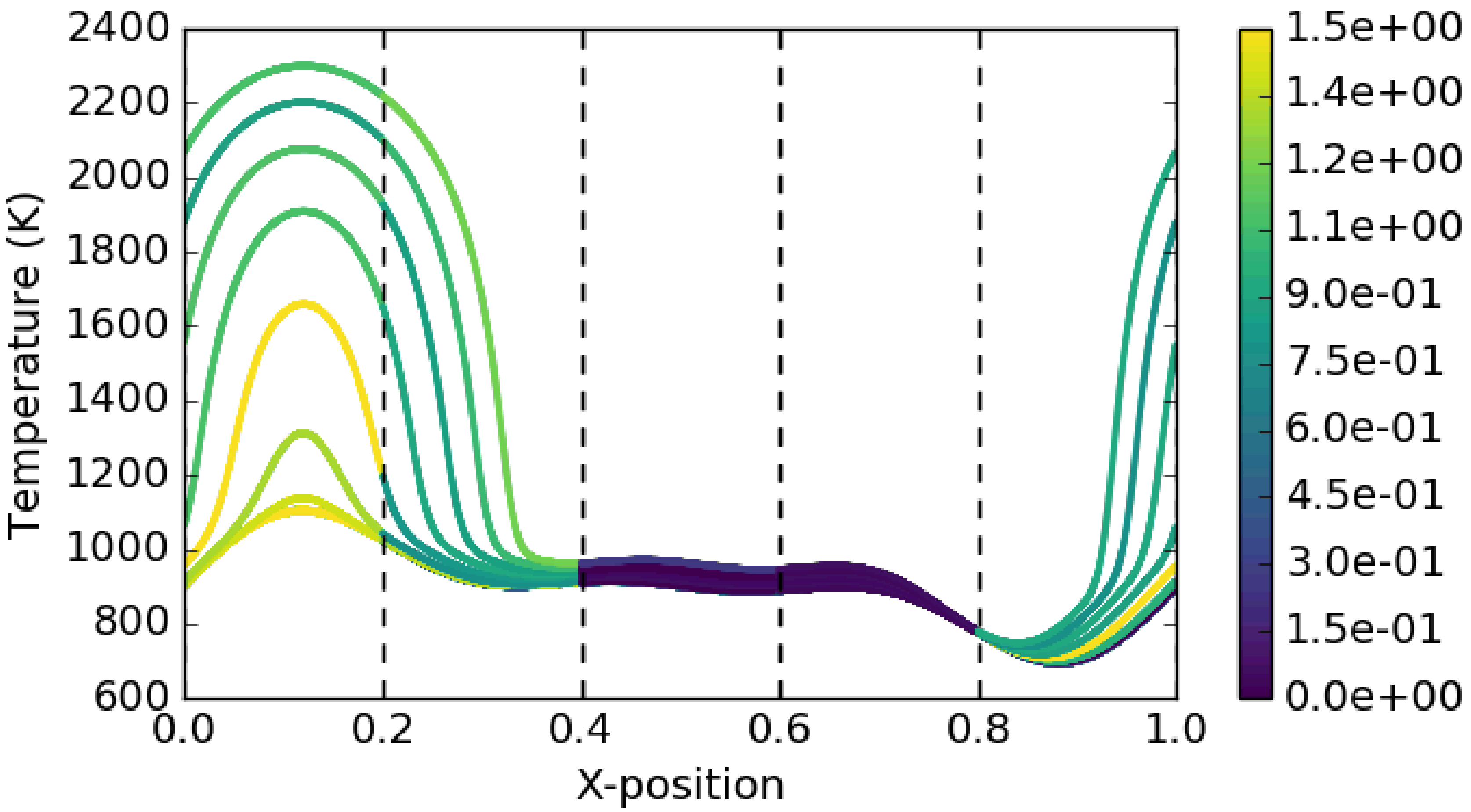


Spiking in both M1 and M2 for Region 1

Auto-Ignition Results (cont.)

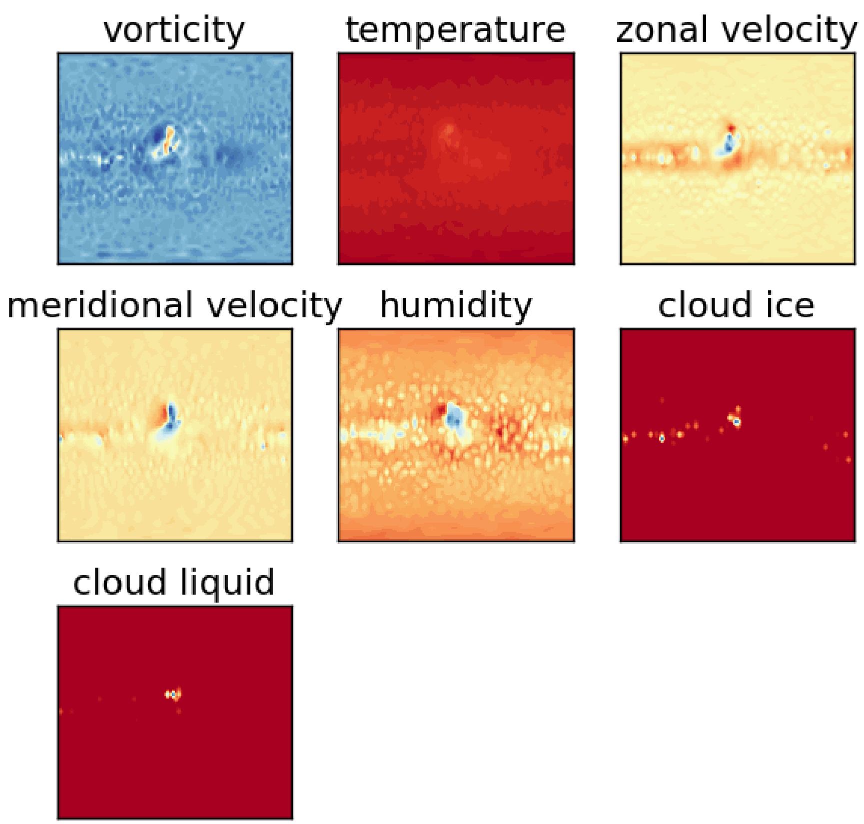
The color gradient shows the M2 metric applied to the temperature profiles.

The M2 values for Region 1 are continually high.

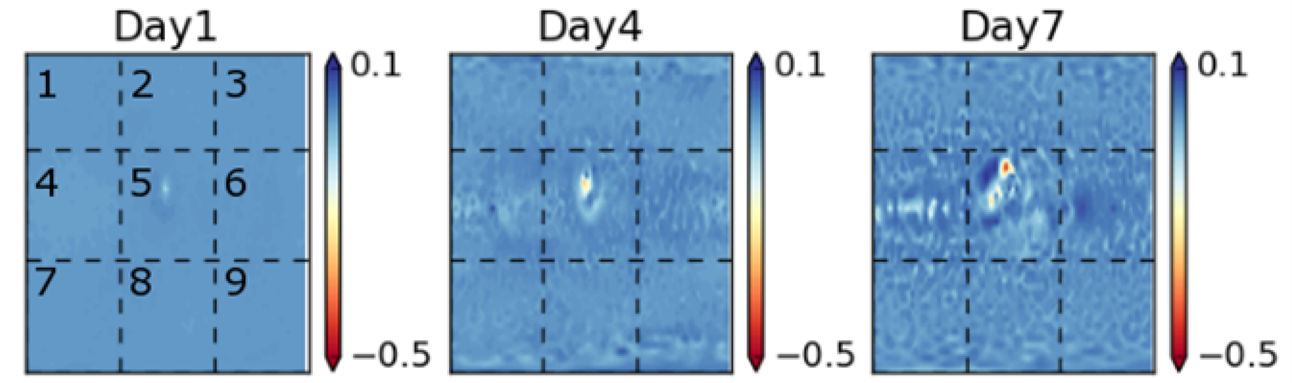


Climate Results

State variable contours



Vorticity over time

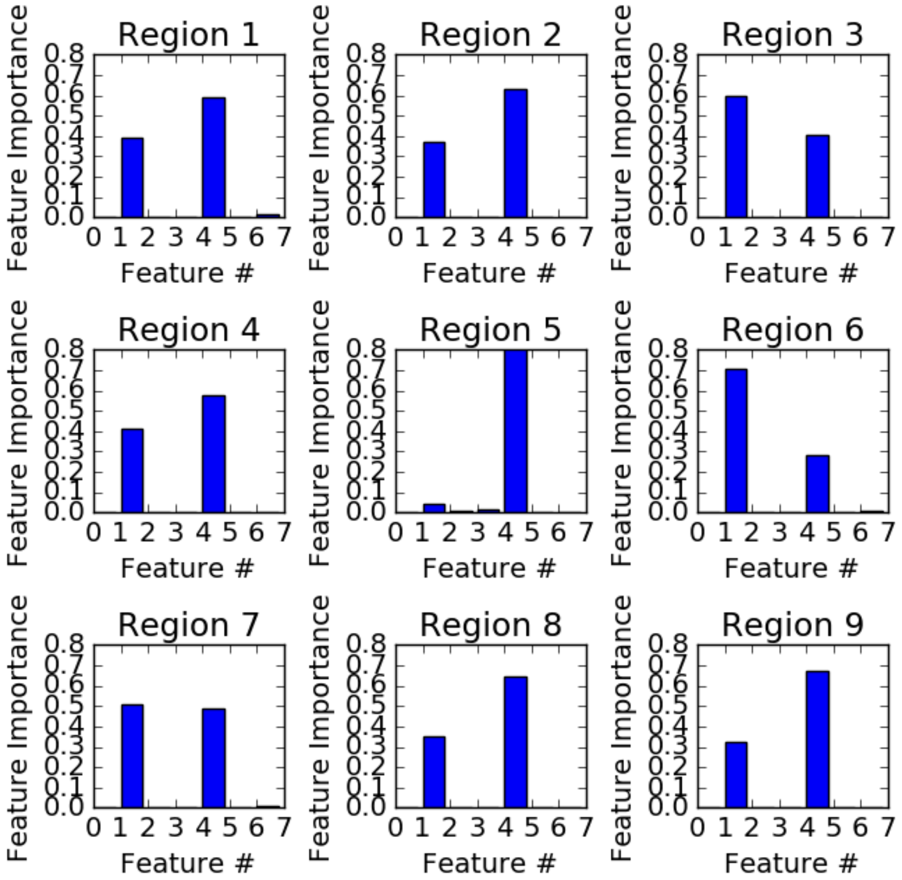


U.S. DEPARTMENT OF ENERGY | Office of Science

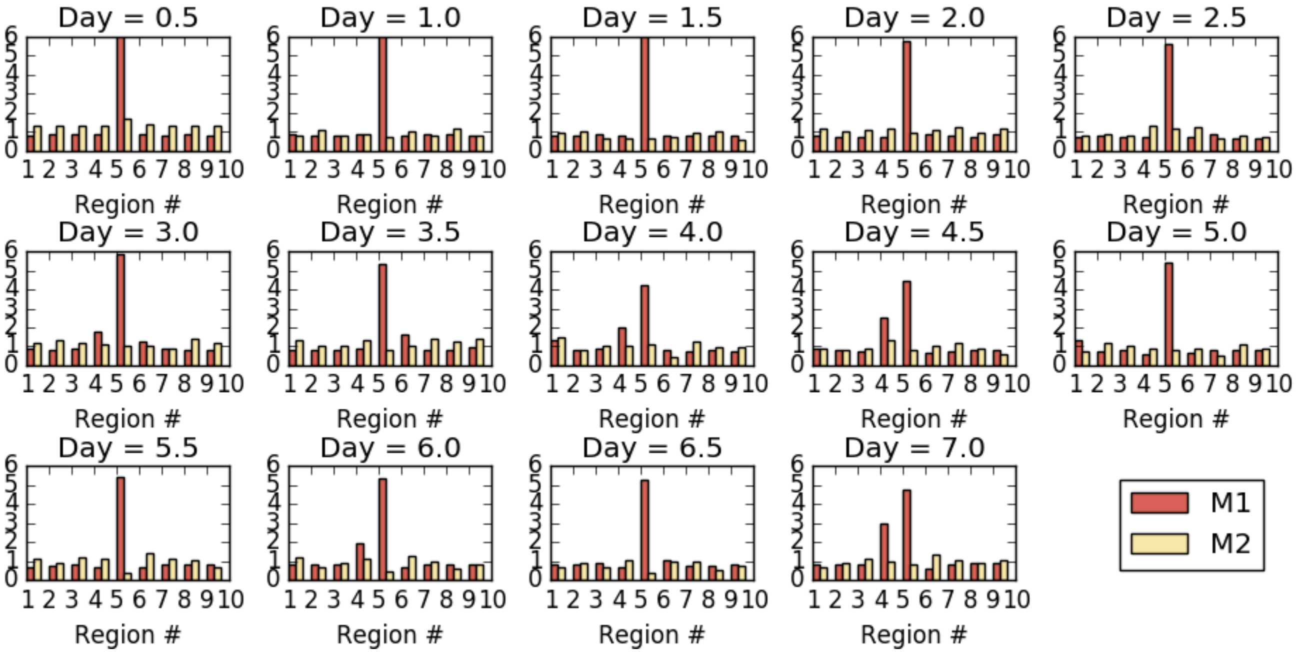
Sandia National Laboratories

Stony Brook University

Climate Results (cont.)



Climate Results show similar effects, with the cyclone being detected spatially in the center of the domain.



U.S. DEPARTMENT OF ENERGY | Office of Science

Sandia National Laboratories Stony Brook University

Summary

- Generated rich test cases within auto-ignition and climate modeling through tight collaboration with our domain experts
- Established vehicle for In-Situ machine learning tests on actual scientific simulations using real hardware
  - Domain experts/developers actively engaged in making this possible
  - S3D is widely used, increasing the potential applicability of this research
- Performed preliminary experimentation in both domains which led to our creation of a new event detection algorithm
  - Preliminary results show great promise
  - Many areas for innovation

Next Steps

- Explore more anomaly detection algorithms
- Explore FIM comparison (distance operators, pdf generation, etc.)
- Begin in-situ experiments
- Explore integration into CAM 5

Publications/Presentations

- “Using Feature Importance Metrics to Detect Events of Interest in Scientific Computing Applications.” Submitted to KDD, 2017

- “In-Situ Machine Learning for Intelligent Data Capture on Exascale Platforms.” Poster presentation at the 2017 Energy and Climate Executive Advisory Board Meeting