EVENT DETECTION IN MULTI-VARIATE SCIENTIFIC SIMULATIONS USING FEATURE ANOMALY METRICS

SAND2018-2210C

Konduri Aditya[1], H. Kolla[1], J. Ling[2], W. P. Kegelmeyer[1], T. Shead[1], D. M. Dunlavy[1], and W. L. Davis[1]

[1] Sandia National Laboratories, Lievrmore, CA, USA

[2]Citrine Informatics, Redwood City, CA, USA

SIAM Conference on Parallel Processing for Scientific Computing 2018, Tokyo, Japan

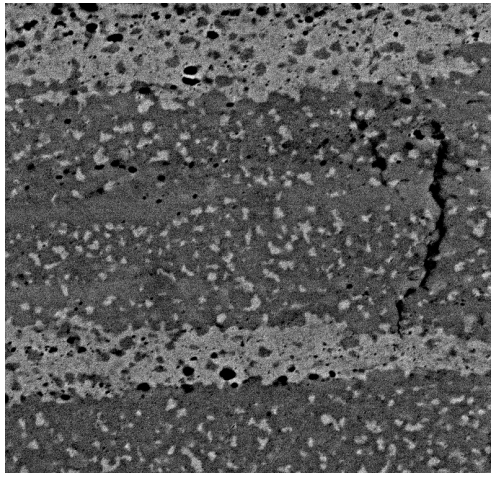
Funding: US Department of Energy Advanced Scientific Computing Research, FWP16-019471

Anomalous or extreme events

Cyclones in weather

source: NOAA

Crack propagation in materials

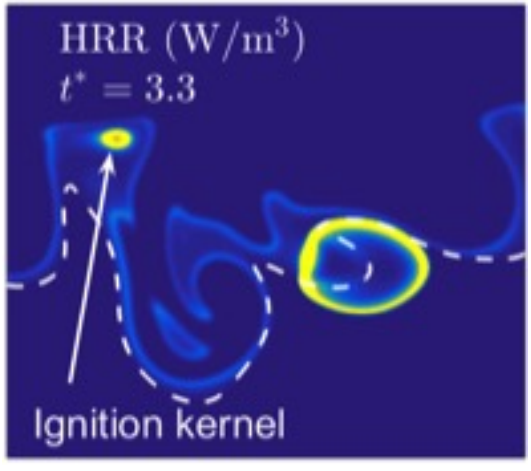


source: K. Satya Prasad

Disease spread in epidemiology

source: Alemi et al. (2015)

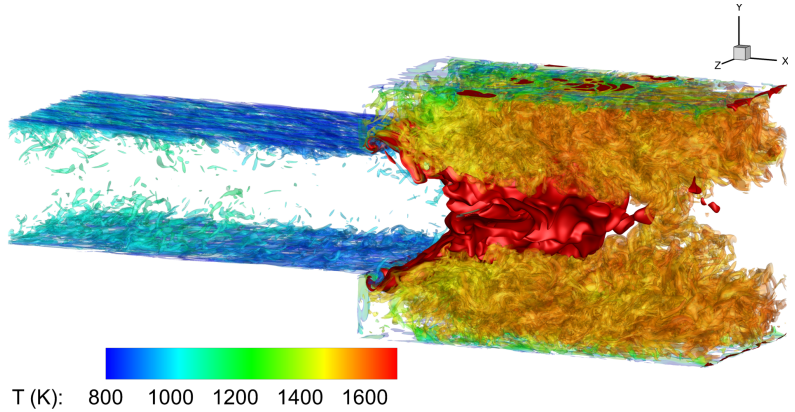
Auto-ignition in combustion



source: Krisman (2016)

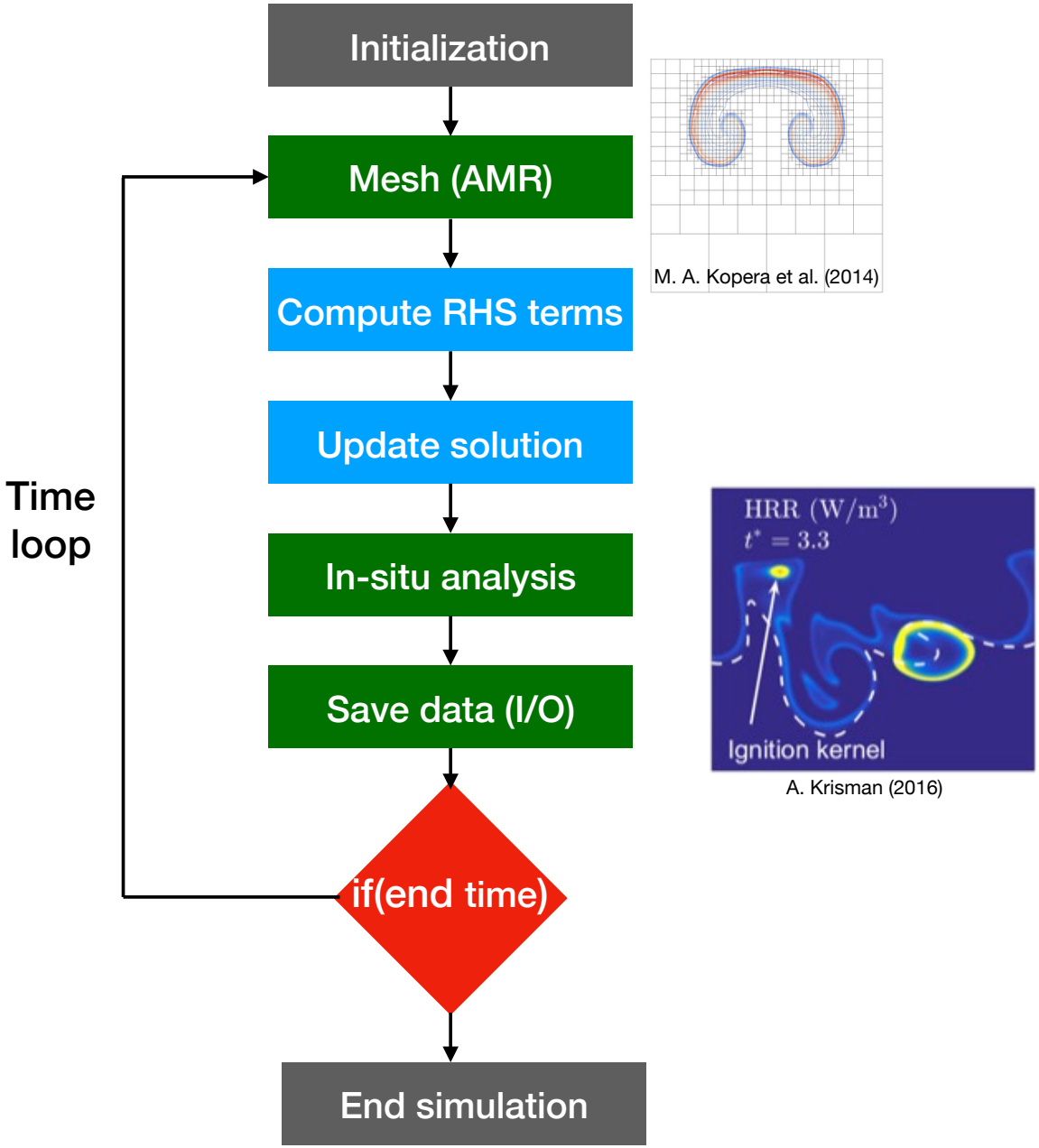
Computational approach

Combustion phenomena



- Governed by highly non-linear PDEs: multi-scale phenomena
- Multi-variate data: $\sim 10 - 100$ variables
- Direct numerical simulations: resolve all the scales in space and time
- Massively parallel solvers (e.g. S3D Chen at al. (CSD 2009))
  - computationally expensive (tens millions of CPU hours)
  - large amount of data ($\sim$ 100 TB)
- **Exascale: need efficient workflows to compute, store and analyze data**

Simulation workflow

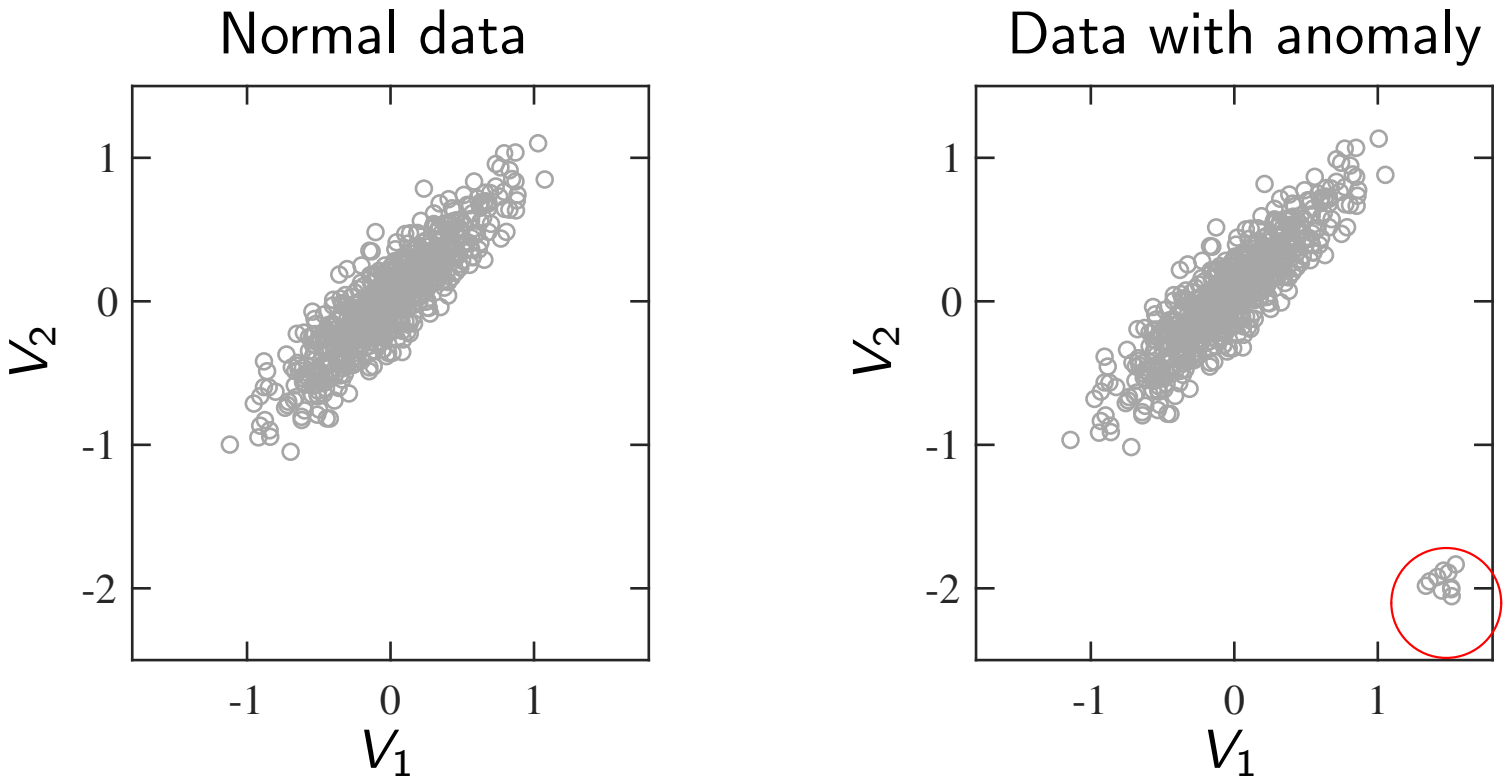


M. A. Kopera et al. (2014)

A. Krisman (2016)

Idea

Bivariate dataset

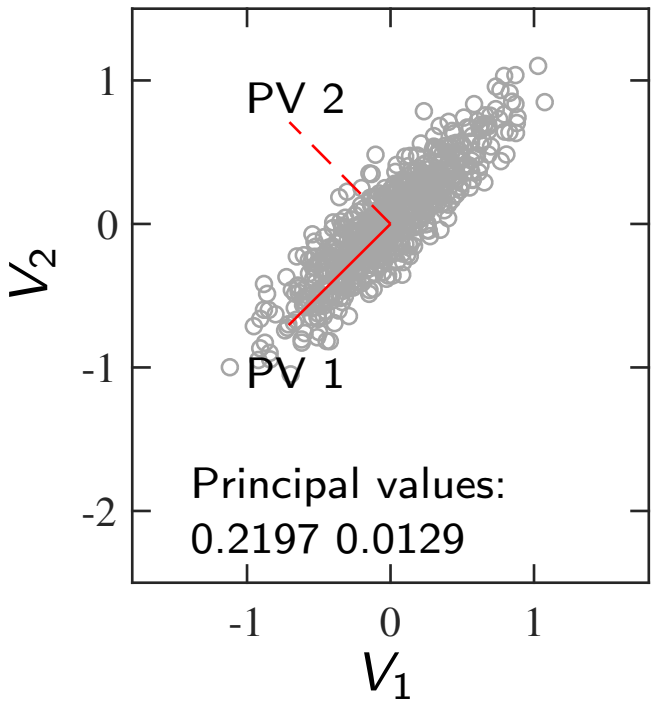


- Characterize the data distribution
- Principal component analysis: principal values and vectors
- Any change in distribution may result in
  - change in the magnitude of principal values
  - change in the orientation of principal vectors

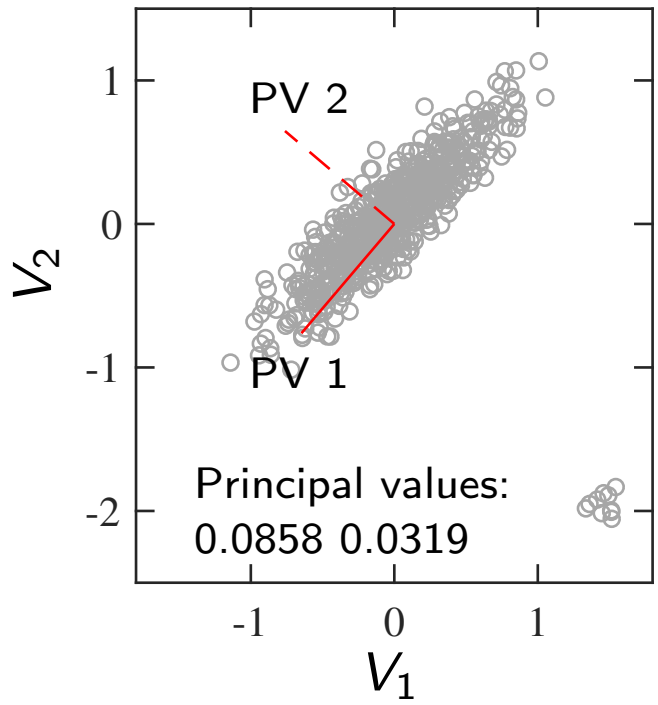
Principal component analysis

- Compute the co-variance matrix (second order joint moment)
- Perform Eigen decomposition to obtain the principal values and vectors

Normal data



Data with anomaly



- First principal vector (PV 1) does not align along the anomalous values
- Mainly captures variance
- Need higher order moments to capture extreme events

Fourth order joint moment

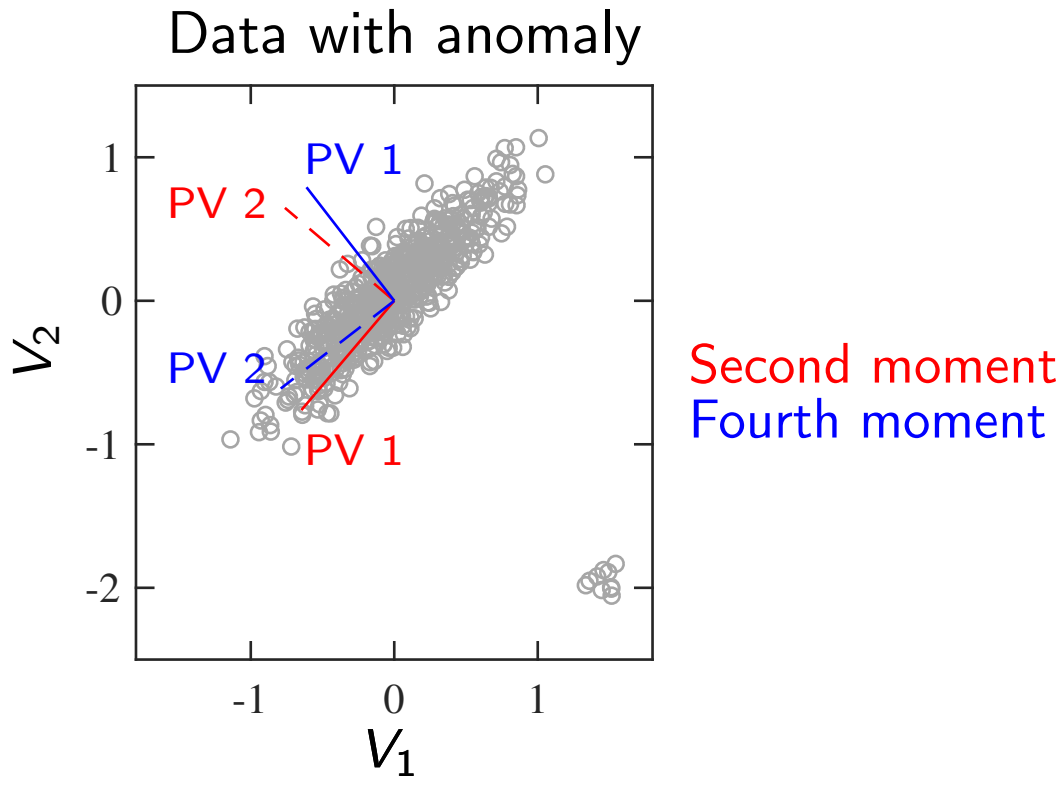
Kurtosis: measure of "either existing outliers (for the sample kurtosis) or propensity to produce outliers (for the kurtosis of a probability distribution)" **(P. H. Westfall, 2014)**

- Compute the fourth joint moment (cumulant tensor, $\mathcal{T}$)

$$\begin{aligned}\mathcal{T} \quad = \quad & \mathbb{E}[v \otimes v \otimes v \otimes v] - \mathbb{E}[v_{i_1} v_{i_2}]\mathbb{E}[v_{i_3} v_{i_4}] \\ & -\mathbb{E}[v_{i_1} v_{i_3}]\mathbb{E}[v_{i_2} v_{i_4}] - \mathbb{E}[v_{i_1} v_{i_4}]\mathbb{E}[v_{i_2} v_{i_3}], \quad 1 \leq i_1 \ldots i_4 \leq k\end{aligned}$$

- Decompose the fourth order symmetric tensor (N_f^4)
  - Canonical polyadic decomposition **(T. G. Kolda, arXiv 2015)**
  - Higher order singular value decomposition **(L. De Lathauwer et al., SIAMJMAA 2015)**
  - Matricize the tensor and perform SVD **(A. Anandkumar et al., JMLR 2014)**
- Obtain principal kurtosis values and vectors

Anomaly detection

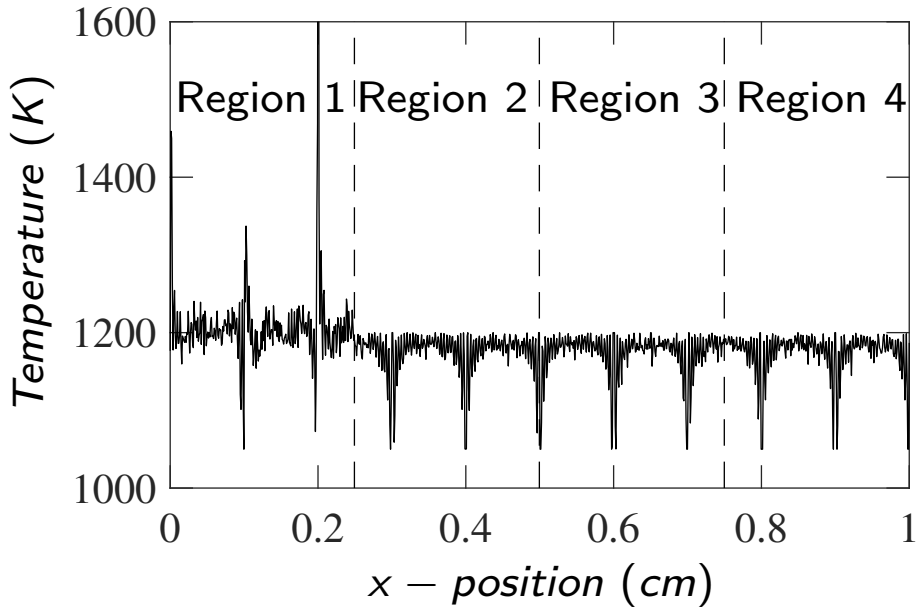


- First principal kurtosis vector aligns in the direction of anomalies
- Can be used to characterize extreme events

Auto-ignition test case

Consider a simple problem with a 1D domain

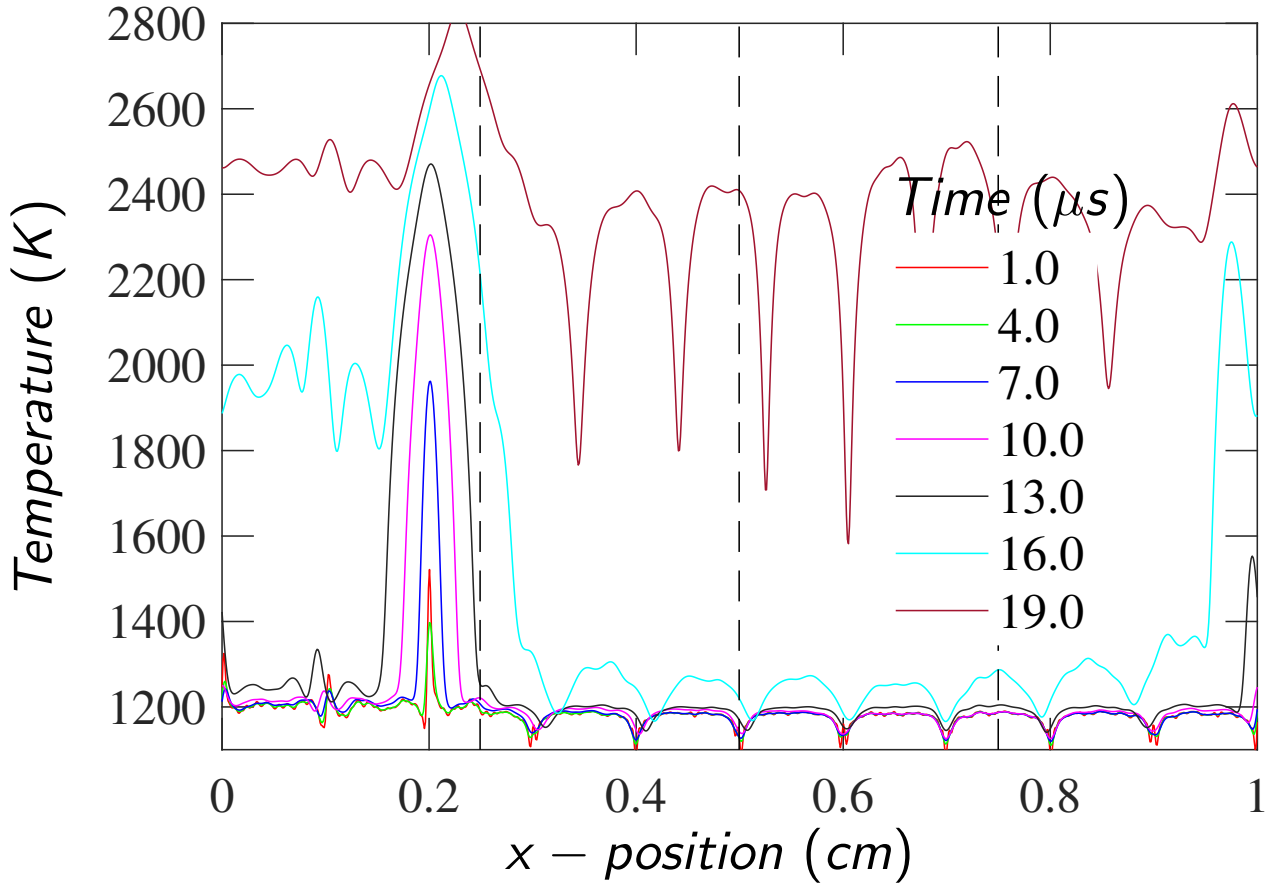
- Initial condition



- Fuel-air composition: $0.6CO + 0.4H_2 + 0.5(O_2 + 3.76N_2)$
- Solver: scalable reacting flow code S3D (**J. H. Chen et al., CSD 2009**)
- Number of subdomains: $N_d = 4$
- Time-steps: $\Delta t = 0.001\mu s$
- Number of checkpoints: $N_t = 20$, interval: $1\mu s$

Auto-ignition test case

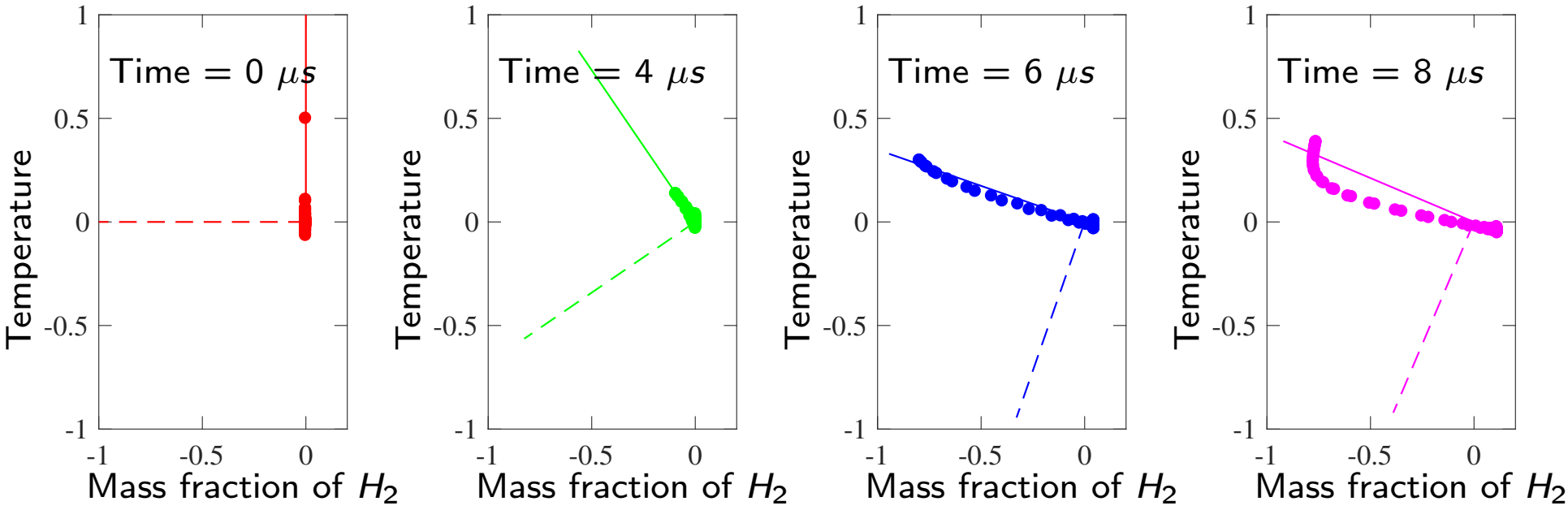
Time evolution of the temperature profiles



- Early ignition occurs in Region 1
- Spatial anomaly in Region 1
- Eventually temporal anomaly in Regions 2, 3, and 4

Principal kurtosis vectors

Time evolution of principal vectors in Region 1



- Initial spread only along temperature
- As ignition event appears, spread along both temperature and H_2
- Principal kurtosis vectors align with ignition event values

Feature moment metric

- Number of features: $N_f = 13$ (12 species + temperature), index i
- Number of subdomains: $N_d = 4$, index j
- Number of time steps: $N_t = 20$, index n
- Project the principal vectors weighted by the principal values onto the features to obtain FMMs

$$F_i^{j,n} = \frac{\sum_{k=1}^{N_f} \lambda_k \left(\hat{e}_i \cdot \hat{v}_k\right)^2}{\sum_{k=1}^{N_f} \lambda_k}$$

- $\hat{e}_i \cdot \hat{v}_k$ is effectively the i-th entry in the k-th vector $\hat{v}_k$
- Property: $\sum_{i=1}^{N_f} F_i^{j,n} = 1, \ \forall\ j, n$

Anomaly metrics

Identify spatial and temporal anomalies

- Statistical signature: distribution of feature moment metrics changes
- Hellinger distance: a symmetric measure of difference between two discrete distributions P and Q

$$D_{PQ} = \frac{1}{\sqrt{2}} \sqrt{\sum_i (\sqrt{p_i} - \sqrt{q_i})^2}$$

- Spatial metric: compare each FMM distribution with the average

$$M_1^n(j) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{N_f} \left(\sqrt{F_i^{j,n}} - \sqrt{\overline{F}_i^n} \right)^2}$$

- Temporal metric: compare FMM distribution between sucessive time steps

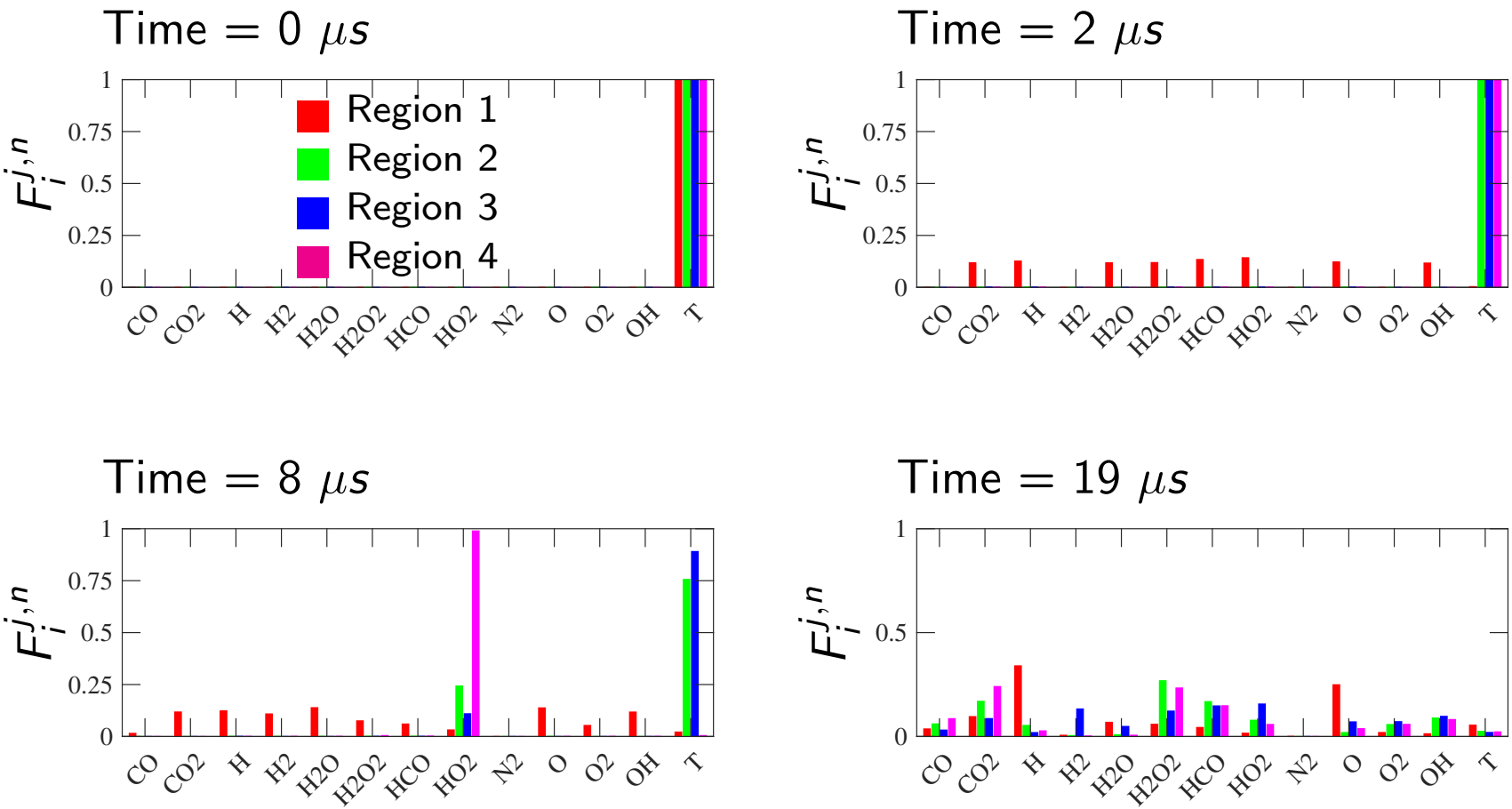
$$M_2^j(n) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{N_f} \left(\sqrt{F_i^{j,n}} - \sqrt{F_i^{j,n-1}} \right)^2}$$

Algorithm

Algorithm 1: Anomaly detection algorithm

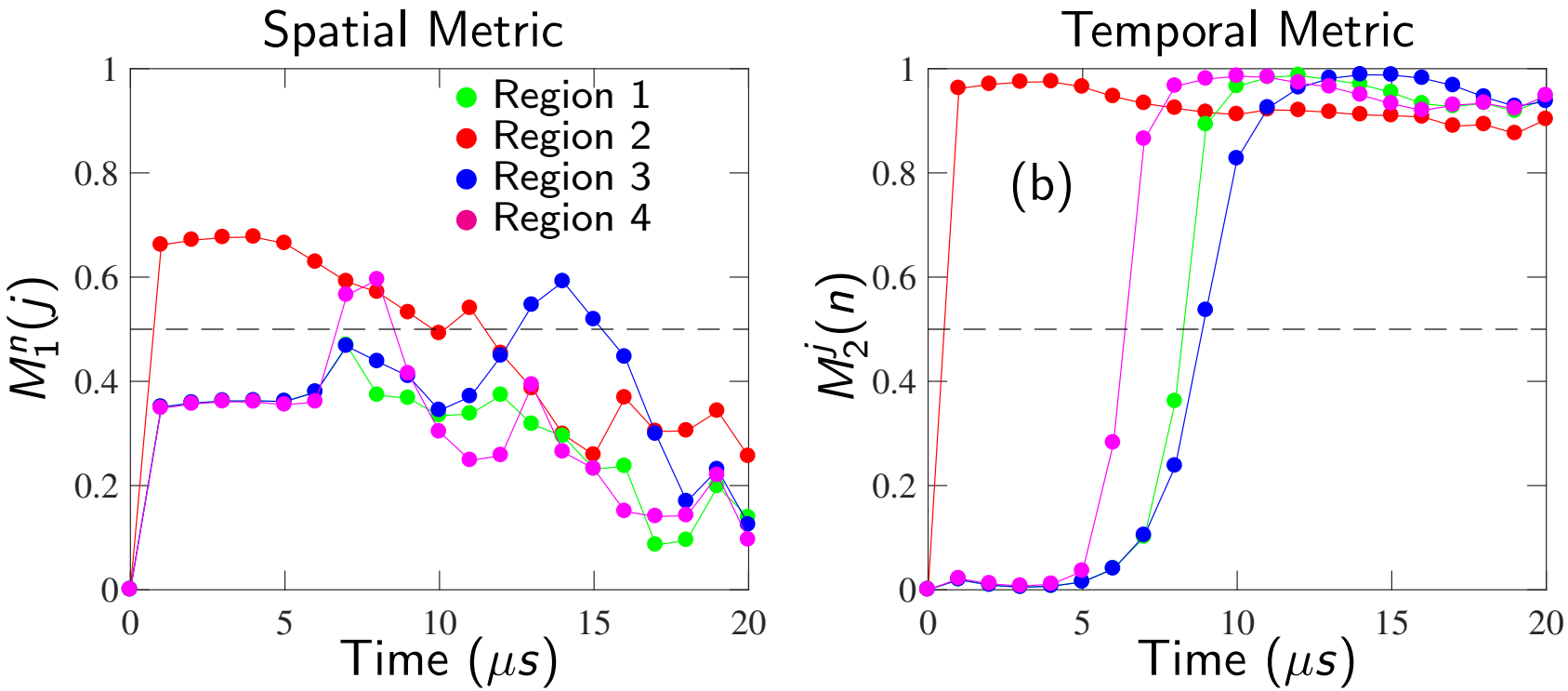
```
   // initialization
1  N_t, N_d ← decompose data;
2  N_f ← select features;
   // time step loop
3  for n ← 1 to N_t do
       // sub-domain loop
4      for j ← 1 to N_d do
5          scale data;
6          T^{j,n} ← compute joint moment tensor;
7          matricize tensor T^{j,n} ;
8          λ_j, v_j ← perform SVD;
9          F_i^{j,n} ← compute feature importance;
10         M_1^n(j), M_2^j(n) ← compute anomaly metrics;
11     end
12     flag anomalous sub-domains;
13 end
```

Feature moment metrics



- FMMs distribute across different features when anomaly occurs

Results



- Dash line: threshold for dectecting anomaly (=0.5)
- Anomalies are detected in space and time

Conclusions

- Proposed an unsupervised anomaly detection algorithm
- Verified the idea using synthetic and auto-ignition data
- Demonstrated the algorithm in a distributed data setting
- Future work:
  - in-situ implementation of the algorithm into the massively parallel direct numerical simulation solver (S3D)
  - evaluate scalability
  - apply the algorithm to detect anomalies in other scientific phenomena

Abstract

In multi-variate multi-physics scientific simulations, anomalous events occur at locations in space-time domain that are hard to predict, for example ignition fronts in combustion. It is often required to identify these events promptly and precisely such that necessary actions may be taken (e.g., triggering in-situ analysis, data checkpoint, mesh refinement), which is challenging in a distributed setting since these events are local in space and/or time. We propose the use of feature anomaly metrics (FAMs) to trigger the detection of such events. Due to tightly coupled physics, anomalies do not always manifest as outliers in individual variables, but as clusters away from the axes in the joint variable space. The FAM quantifies the contribution of each variable to anomalies in the joint variable space based on its alignment with vectors that point towards anomalous clusters. To construct such vectors, we seek a change of basis in a manner analogous to PCA. While PCA yields a change of basis guided by the co-variance matrix, a measure of data spread, we desire a change of basis guided by a measure of data outlierness, co-Kurtosis, which is a symmetric fourth-order tensor. We employ symmetric CP decomposition of the co-Kurtosis tensor to perform a change of basis and construct FAMs. We examine the efficacy of FAMs in identifying anomalous events in synthetic data as well as canonical 1D combustion simulation data.