Machine Learning and Topological Data Analysis: Applications to Pattern Classification in Fluid and Climate Simulations

Vitaliy Kurlin[1], Grzegorz Muszynski[1,2], Michael Wehner[2], Karthik Kashinath[2], Prabhat[2]

1) Computer Science Department, University of Liverpool, United Kingdom
2) Lawrence Berkeley National Laboratory, United States

Big Data Summit, NERSC, Berkeley, CA
July 18, 2018

UNIVERSITY OF LIVERPOOL

Outline of the talk

Team members and goals of the project

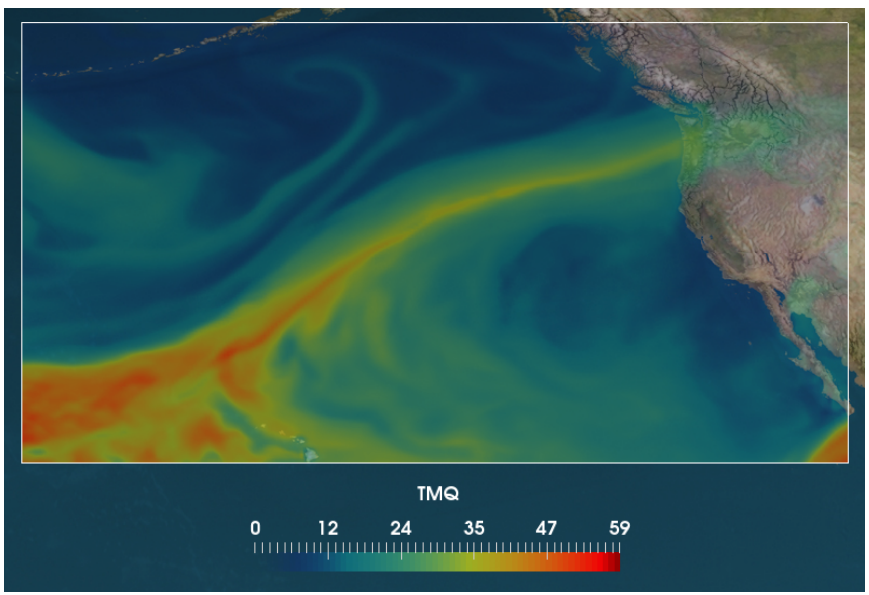
Grzegorz has started his PhD at the University of Liverpool in April 2017, co-supervised by Michael Wehner and Vitaliy Kurlin.

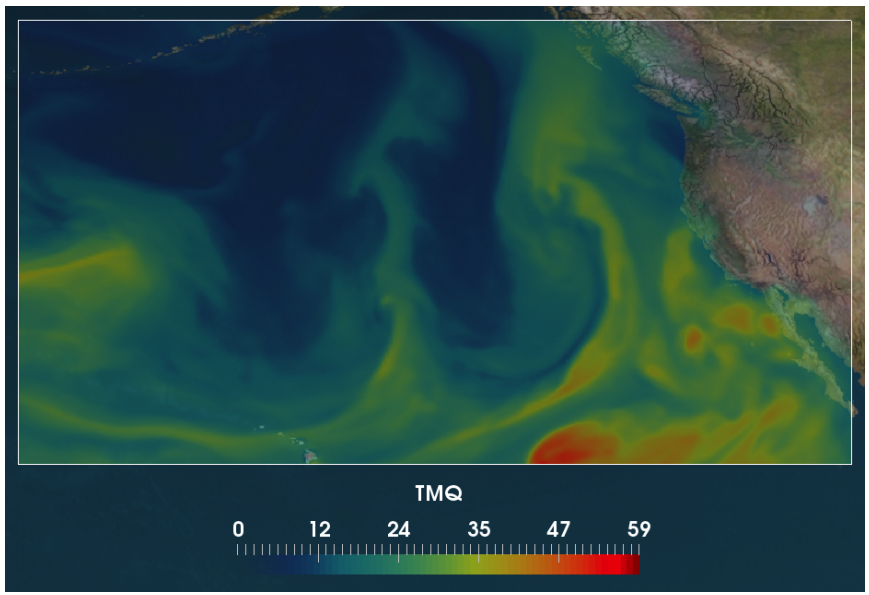
Prabhat and Karthik have substantially contributed to the project.

- Find low-dimensional representations of data that capture non-linear dependencies in climate data.
- Develop topological data analysis methods for detecting and classifying patterns ("shapes") in climate data.
- Design and implement algorithms that can be included into the Toolkit of Extreme Climate Analysis (TECA) used at the Berkeley Lab for distributed and parallel computing.

Science Problem: Motivation

- Understand changes in **extreme weather events** or patterns. An *Atmospheric River* (AR) is a long narrow high-moisture filament. ARs play a key role in the global water-cycle.
- Produce useful **frequency statistics** for climate models based on the number of AR occurrences during a year.

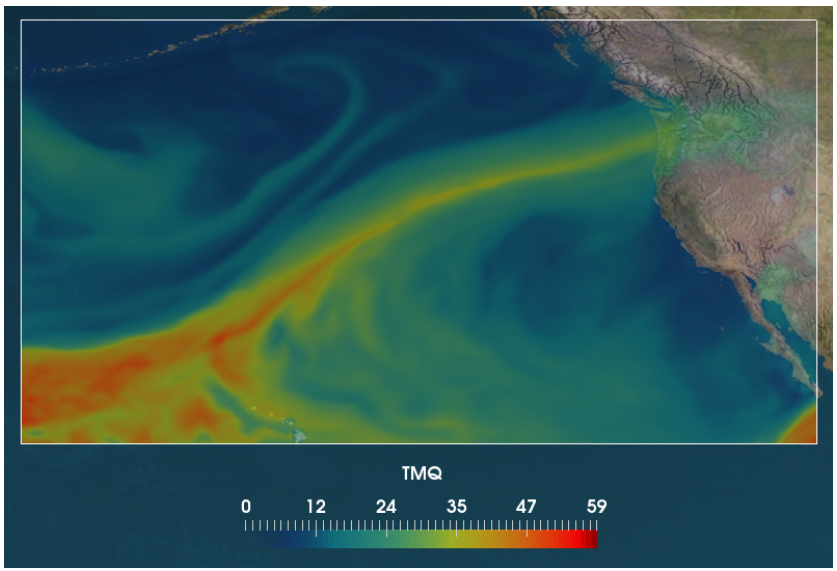


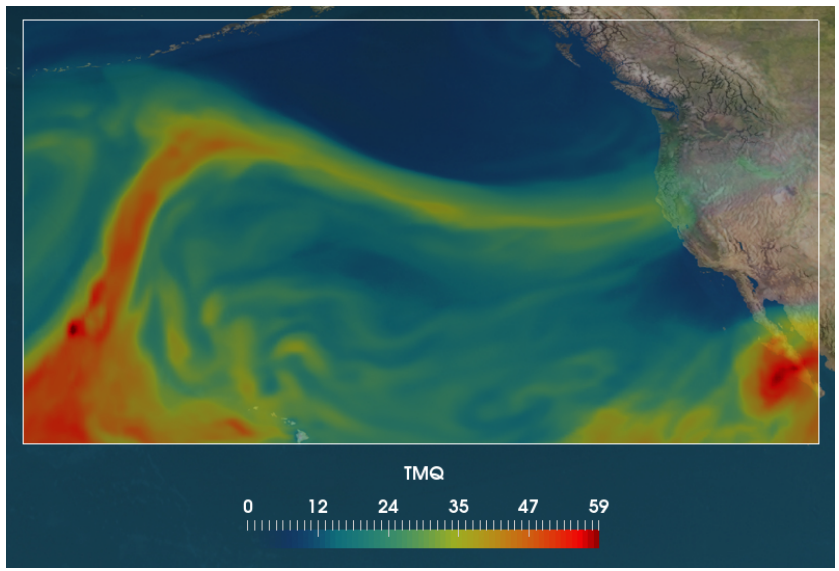


Left: an AR bringing water vapour from the tropics to the western US. **Right**: a **non-AR** event without a filamentary structure. *TMQ* is the Integrated Water Vapor (IWV) in kg/m^2 (mass over area).

Science Problem: Goals

- **Avoid subjective thresholds** on physical quantities such as 20 kg/m^2 of Integrated Water Vapor (IWV).
- Provide a **reliable weather/climate pattern recognition method** that can work without manual tuning of parameters for *different resolutions* of climate models.
- **Identify ARs** with high accuracy and precision.



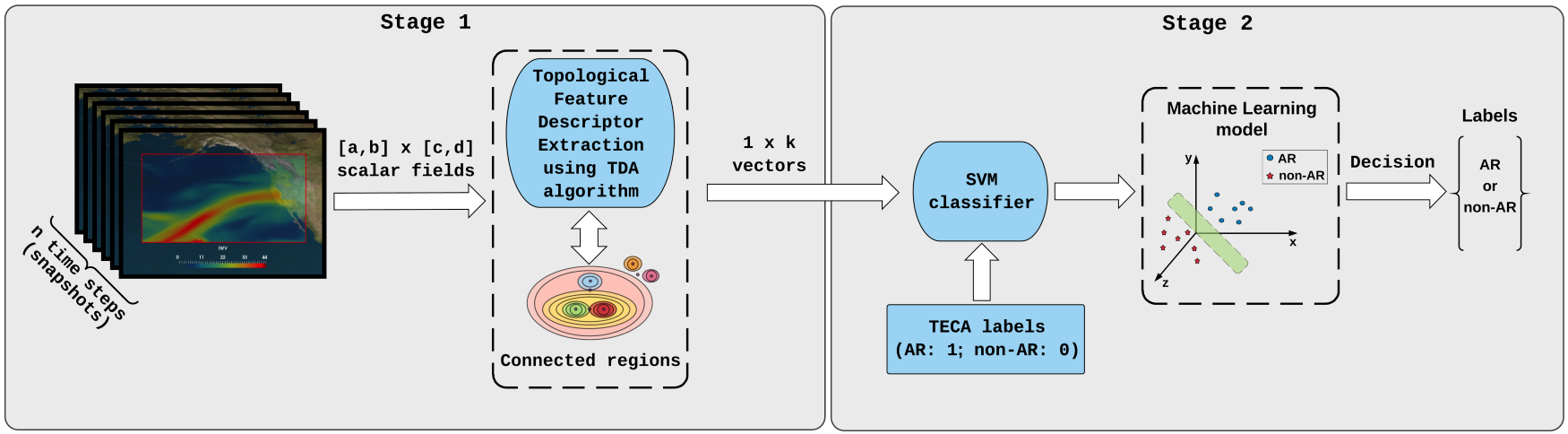


Atmospheric rivers (ARs) can have very different geometric shapes and features (lengths and widths). Topological features of ARs (connectivity, holes) are invariant under continuous deformations.

Method: Pattern Recognition Method

We developed a **two-stage method for AR pattern recognition** based on *Topological Data Analysis*, *i.e.*, TDA algorithm, and a machine learning algorithm, *i.e.*, *Support Vector Machine* (SVM).

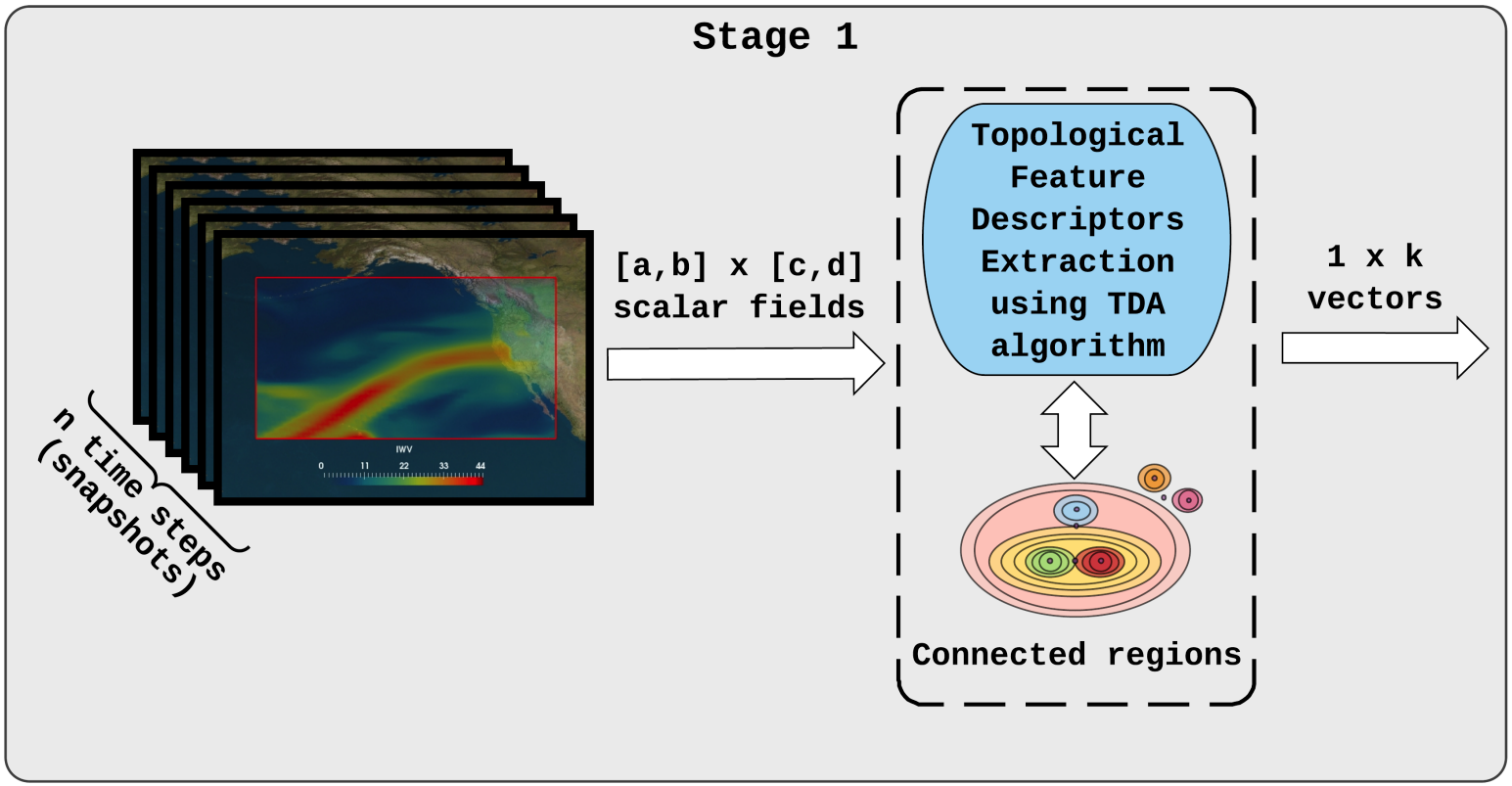
- **Input:** 2D scalar fields (IVW) on a regular grid.
- **Output:** binary labels: AR = 1, non-AR = 0.



The flowchart of two stages of the AR pattern recognition method.

Method: Stage 1 of the proposed AR detection

- **Stage 1**: extract topological descriptors (sizes of connected regions). The TDA algorithm uses the Union-Find data structure and runs in time $\mathcal{O}(m \log m)$ for m grid points.

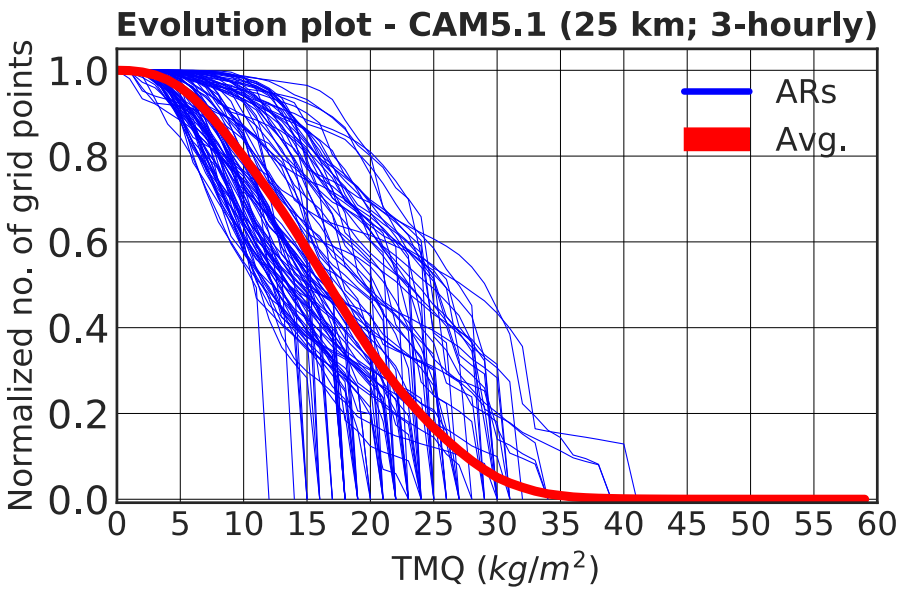


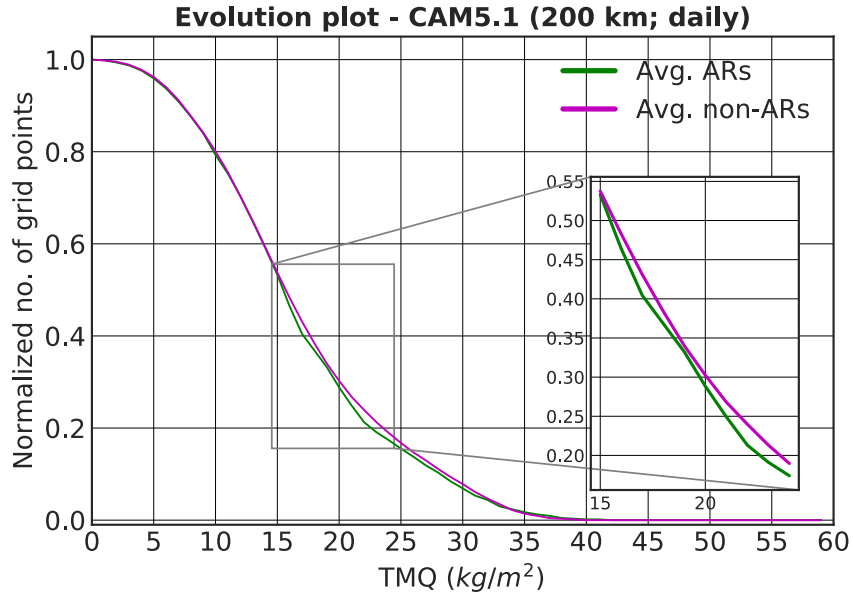
Stage 1: extract topological descriptors from 2D scalar fields.

Method. Stage 1: Topological Data Analysis

The algorithm monitors changes in superlevel sets consisting of all grid points with IVW higher than a variable threshold.

At a critical moment, two locations (*e.g.*, Hawaii and the West Coast) will be covered by a single connected high moisture region.

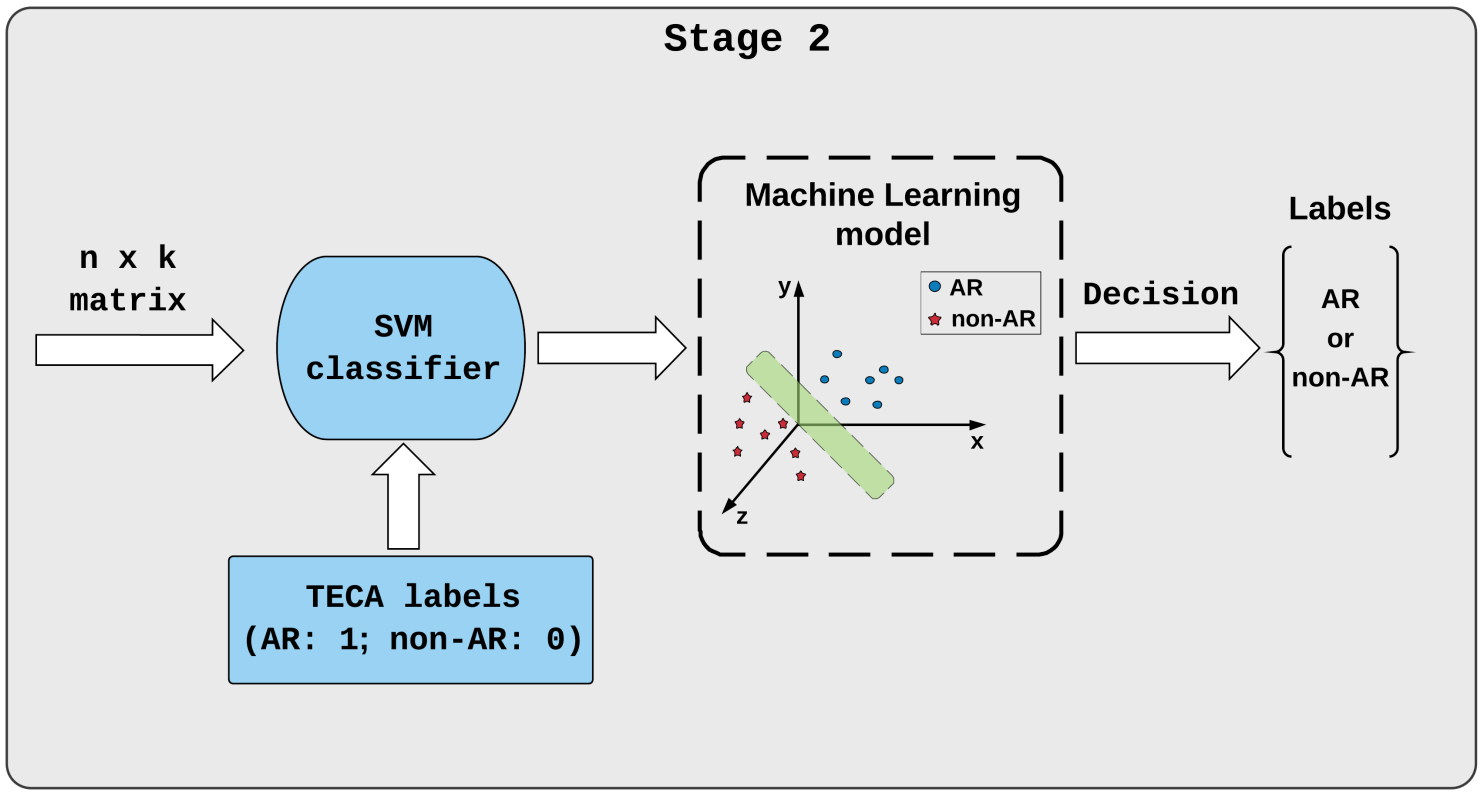




Sizes of superlevel sets vs thresholds. **Left**: 100 random ARs; **Right**: Averaged and normalized descriptors for the full dataset.

Method. Stage 2: Support Vector Machines

- Vectorize topological descriptors, train SVM on TECA labels.
- Exhaustively search for best hyper-parameters in a grid, *i.e.* loose and fine grid searching approaches are applied.
- Perform a cross-validation classification.



Stage 2: classify ARs using topological features of 2D scalar fields and labels from *TECA* (the Toolkit for Extreme Climate Analysis).

Method: Implementation

C++ and Python code:

- We have implemented TDA algorithm in C++ inside LBNL's parallel toolkit *TECA* (the Toolkit for Extreme Climate Analysis) using distributed data parallelism and map-reduce.
- We can clean and optimize the current C++ implementation of the TDA algorithm if needed.
- We have used Support Vector Machine classifier (SVM) from the *scikit-learn* machine learning package in Python.

Method: Implementation's Performance

Runs on Cori:

- Runs on 2-15 "Haswell" nodes (32 cores each).
- Loads up to several dozens of GB.
- Execution time of the algorithm:
  - TDA stage: 6 minutes - 15 minutes.
  - SVM stage (exhaustive grid search for hyper-parameters tuning): 40 minutes - 4.5 hours.

Performance or Scaling issues observed:

- SVM's skitlearn implementation does not perform as expected, i.e., it only provides parallelism for hyper-parameters tunning.
- Need to use Intel Data Analytics Acceleration Library (Intel DAAL) for the SVM.

Results: Science Results

- New way of analyzing and understanding the behaviour of weather patterns, in particular Atmospheric Rivers (ARs).
- Novel technique for assessment of climate model by using TDA & machine learning (SVM) framework.
- The method has no thresholds and works for different resolutions of climate data in the table below.

Climate Model	Period	Temporal Resolution	Spatial Resolution
CAM5.1 (historical run)	1979-2005	3-hourly and daily	25 km
CAM5.1 (historical run)	1979-2005	3-hourly and daily	100 km
CAM5.1 (historical run)	1979-2005	3-hourly and daily	200 km
MERRA-2 (reanalysis product)	1980-2017	3-hourly	50 km

These climate datasets were used to test the method.

Classification Accuracy on the 3-hourly data

Climate Dataset	**#ARs snapshots**	**#Non-ARs snapshots**	**Train ACC.**	**Test ACC.**
CAM5.1 (25km)	6838	6848	83%	83%
CAM5.1 (100km)	7182	7581	77%	77%
CAM5.1 (200km)	3914	3914	90%	90%

Table 1: Classification accuracy score for 3-hourly temporal resolution of the Community Atmosphere Model, Version 5.1 with three different spatial resolutions.

Classification Accuracy on the daily data

Climate Dataset	**#ARs snapshots**	**#Non-ARs snapshots**	**Train ACC.**	**Test ACC.**
CAM5.1 (25km)	624	624	78%	82%
CAM5.1 (100km)	700	700	85%	84%
CAM5.1 (200km)	397	397	89%	91%

Table 2: Classification accuracy score for daily temporal resolution of the Community Atmosphere Model, Version 5.1 (CAM5.1) with three different spatial resolutions.

Classification Accuracy on the larger data

Climate Dataset	**#ARs snapshots**	**#Non-ARs snapshots**	**Train ACC.**	**Test ACC.**
MERRA2 (50km)	13294	13434	80%	80%

Table 3: Classification accuracy score for 3-hourly temporal resolution of the Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA2) with 50 km spatial resolution.

Results: Conclusions

- Topological algorithm reduced the feature extraction time to **couple of minutes** in comparison with training of CNN networks (*i.e.*, hours or days);
- New technique for **feature extraction of weather patterns** in climate data.
- Improved classification accuracy and precision up to **91%** and **0.97**, respectively.

Outcomes: Papers

- Shields, C. A., Rutz, J. J., Leung, L.-Y., Ralph, F. M., **Wehner, M.**, Kawzenuk, B., Lora, J. M., McClenny, E., Osborne, T., Payne, A. E., Ullrich, P., Gershunov, A., Goldenson, N., Guan, B., Qian, Y., Ramos, A. M., Sarangi, C., Sellars, S., Gorodetskaya, I., **Kashinath, K.**, **Kurlin, V.**, Mahoney, K., **Muszynski, G.**, Pierce, R., Subramanian, A. C., Tome, R., Waliser, D., Walton, D., Wick, G., Wilson, A., Lavers, D., **Prabhat**, Collow, A., Krishnan, H., Magnusdottir, G., and Nguyen, P.:
  Atmospheric River Tracking Method Intercomparison Project (ARTMIP): Project Goals and Experimental Design,
  Geoscientific Model Development, **published** on 20 June 2018,
  `https://doi.org/10.5194/gmd-2017-295`
- **Muszynski, G.**, **Kashinath, K.**, **Kurlin, V.**, **Wehner, M.**, **Prabhat**:
  Topological Data Analysis and Machine Learning for Recognizing Atmospheric River Patterns in Large Climate Datasets,
  Geoscientific Model Development, **under review** since February 2018,
  `https://www.geosci-model-dev-discuss.net/gmd-2018-53/`.
- **Muszynski, G.**, **Kurlin, V.**, **Morozov, D.**, **Kashinath, K.**, **Wehner, M.**, **Prabhat**:
  Topological Methods for Pattern Detection in Climate Data,
  a book chapter for Wiley & Sons, **under review** since March 2018,
  invited by Huang Thomas, Jet Propulsion Laboratory at Caltech.

Outcomes: Talks

- Talk at *Learning Algorithms* session at **the British Colloquium for Theoretical Computer Science** at Royal Holloway, University of London, the United Kingdom, 26 March 2018.
- Talk at *Big data and machine learning in geosciences* session at **the European Geosciences Union General Assembly (EGU)**, Vienna, Austria, 9 April 2018.
- Invited talk at *Algorithms, Computational Geometry and Topology* seminar at **the Institute of Science and Technology Austria (IST Austria)**, Vienna, Austria, 11 April 2018, **invited by Wagner Hubert, Edelsbrunner Group**.
- Talk at **Postgraduate Workshop** at the Department of Computer Science, University of Liverpool, the United Kingdom, 1 May 2018.
- Talk at **Applied and Computational Topology Meeting** organized by Applied Algebraic Topology network, University of Southampton, Southampton, the United Kingdom, 30 April 2018.
- Lighting talk at **Data-Driven Modelling of Complex Systems** at **The Alan Turing Institute**, London, the United Kingdom, 8 May 2018.
- Talk at **2018 International Atmospheric Rivers Conference**, SCRIPPS Institute of Oceanography at University of California San Diego, the United States, 26 June 2018 (given by Karthik Kashinath).

Outcomes: Posters

- Poster at Spring School on **Applied and Computational Algebraic Topology** at Hausdorff Research Institute for Mathematics, Bonn, Germany, 24-28 April 2017.
- Poster at **the Conference on Applied and Computational Algebraic Topology** at Hausdorff Research Institute for Mathematics, Bonn, Germany, 2-6 May 2017.
- Poster at **the Computing Sciences Summer Student poster session** at Lawrence Berkeley Lab, Berkeley, California, United States, 10 August 2017.
- Poster at *the An Object-Oriented View of Atmospheric Science: Feature Detection and Characterization in Big Data*, **the American Geophysical Union Fall meeting** in New Orleans, United States, 11 December 2017.
- Poster at **2nd ARTMIP | Atmospheric River Tracking Method Intercomparison Project workshop**, Gaithersburg, Maryland, the United States, 23-24 April 2018 (presented by Vitaliy Kurlin).
- Poster at **Data-Driven Modelling of Complex Systems** at **The Alan Turing Institute**, London, the United Kingdom, 8-10 May 2018.

Outcomes: Community Outreach

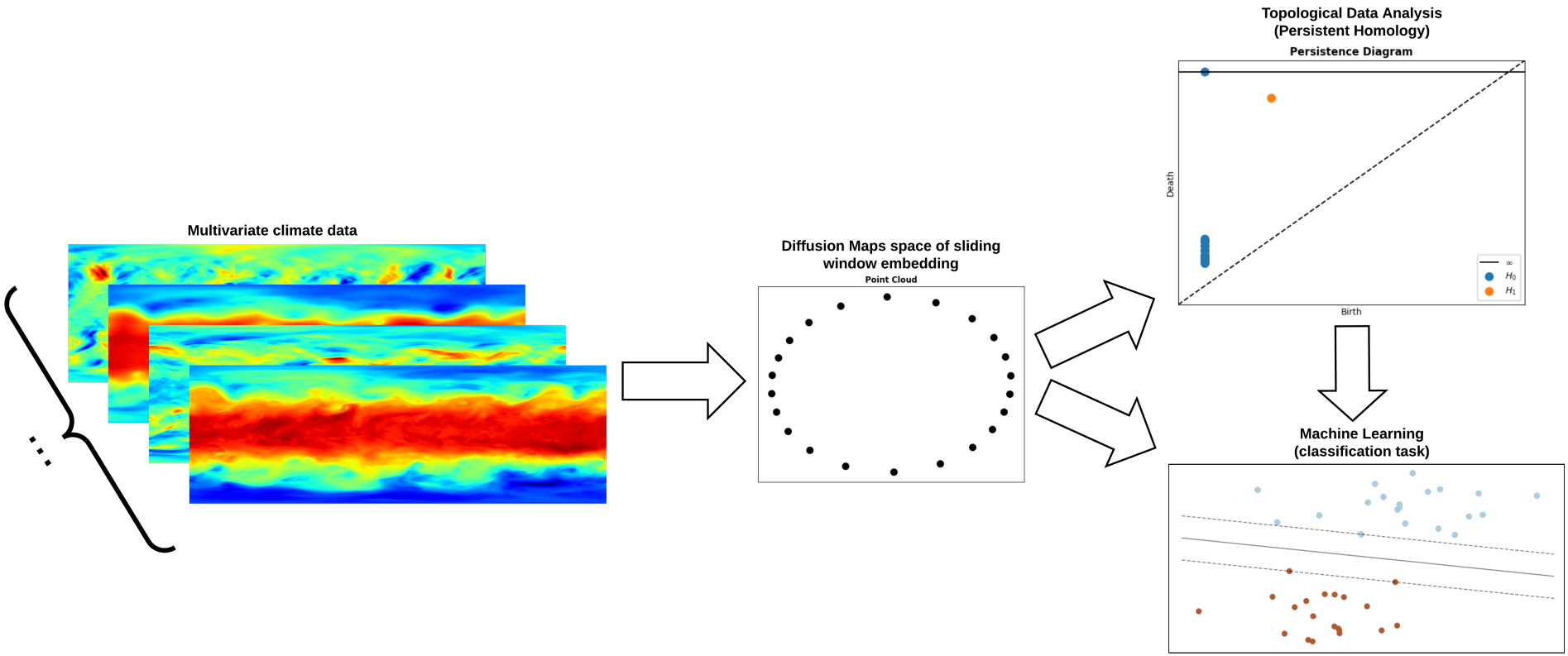
- Discussions with **a world-leading practitioners of applied topology**, like **Professor Gunnar Carlsson** (co-founder of Ayasdi company) and **Professor John Harer** (Duke).
- Meeting with **one of the world-leading groups in algorithms, computational geometry and topology** (*i.e.*, **Edelsbrunner group**) at Institute of Science and Technology Austria (IST Austria).

Future Work: more science problems and motivations

- Detect persistent patterns (e.g., swirling vortices) in fluid flow simulation data.
- Detect climate patterns (e.g., atmospheric blocking - high pressure pattern) in climate simulation and reanalysis data.
- Design and develop a detection method of patterns in fluid flow simulation and climate data.
- We plan to combine dynamical systems theory, manifold learning and topological data analysis methods.

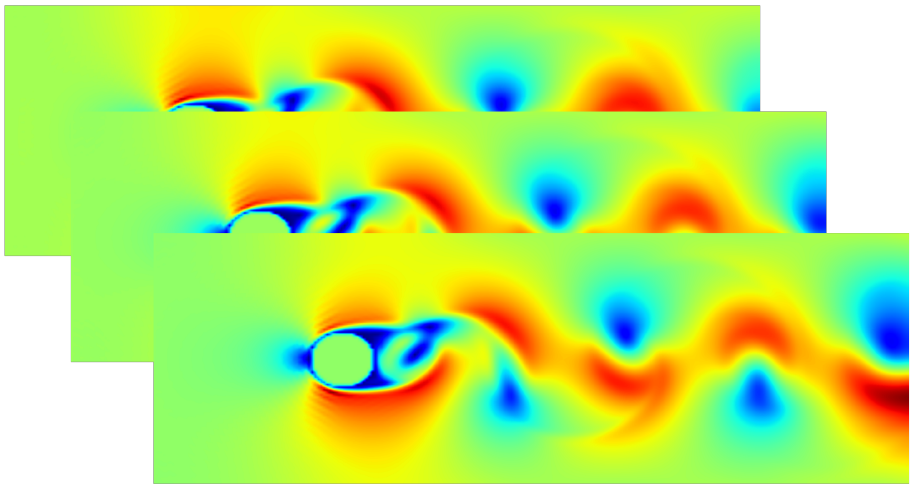
Future Work: new framework

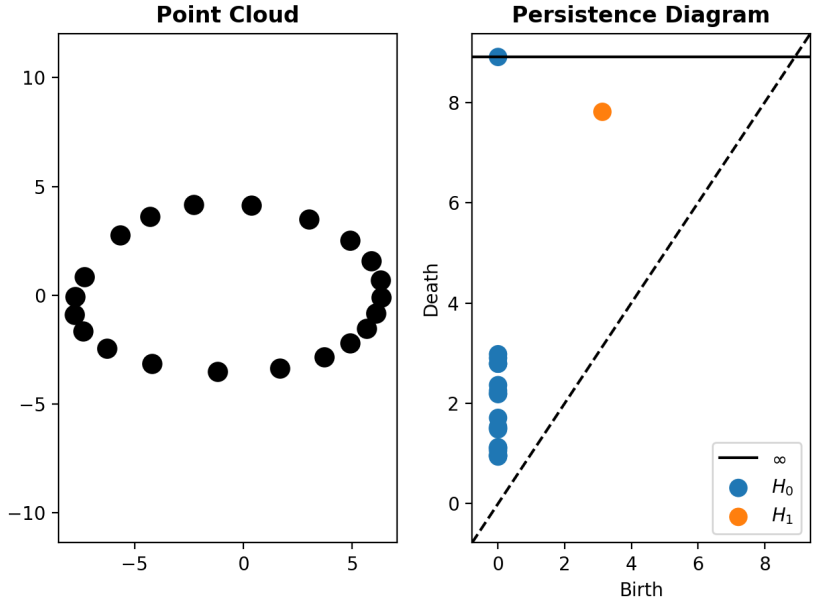
New framework combines: Taken's time-delay coordinate embedding, diffusion maps dimensionality reduction algorithm and topological data analysis (i.e. persistent homology).

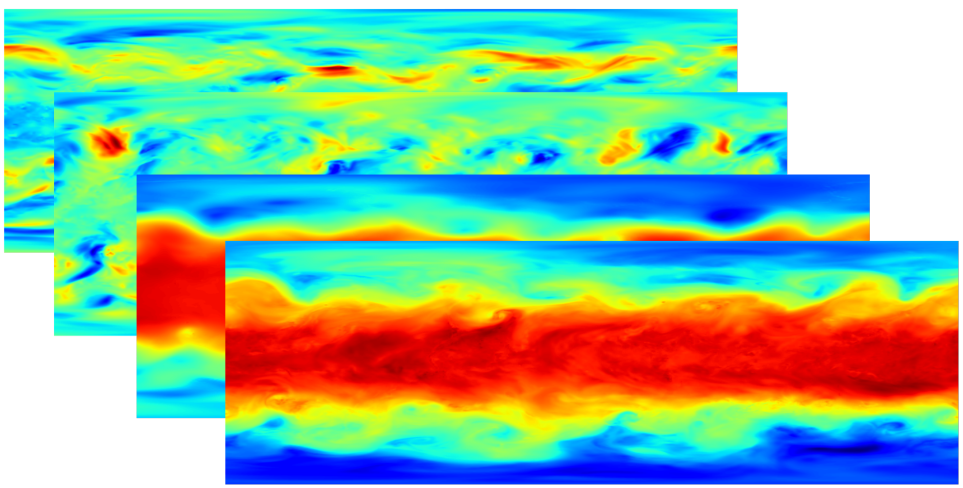


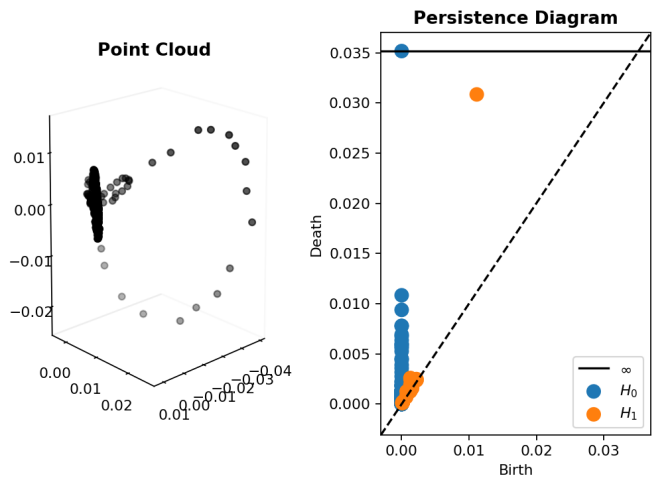
The flowchart above illustrates the approach to topological pattern detection in fluid or climate simulation data.

Future Work: a proof-of-concept









Preliminary results for fluid flow (upper row) and climate data (lower row).

Future Work: potential performance

- Computing Taken's time-delay (sliding window) coordinate embedding: $\mathcal{O}(mn)$ for n timesteps and window size of m.
- Computing sparse diffusion maps dimensionality reduction algorithm: $\mathcal{O}(k)$ for k elements in matrix.
- Computing topological data analysis, i.e. persistent homology using Ripser[1] that outperforms other implementations by a factor of more than 40 in computation time and a factor of more than 15 in memory efficiency.

[1] Ulrich Bauer, Ripser (C++), https://github.com/Ripser/ripser

Thank you!

Vitaliy Kurlin and Grzegorz Muszynski would like to thank **Intel** for funding the IPCC at Liverpool and **NERSC** for computational resources!

We also thank **Prabhat**, **Karthik Kashinath**, and **Michael Wehner** and other NERSC staff!