Topological Data Analysis and Machine Learning for Detecting Atmospheric River Patterns in Climate Data

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

1) Department of Computer Science, University of Liverpool, UK
2) National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory, Berkeley, US

2018 International Atmospheric Rivers Conference (IARC)

26 June 2018

Q: Can we automatically identify weather patterns in a climate model output?

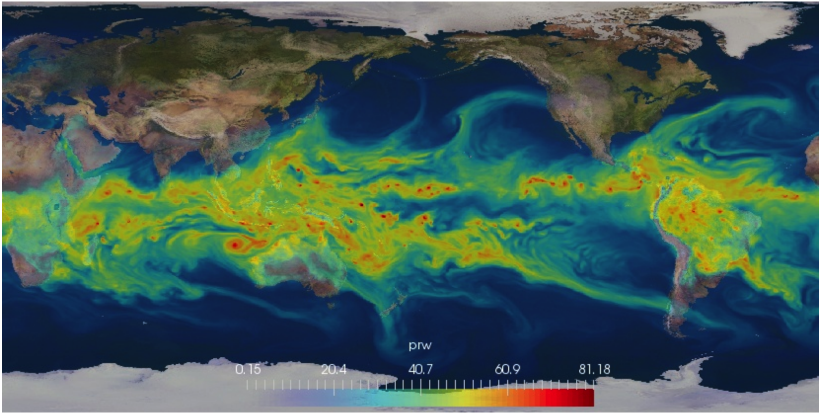


Figure 1: An example of climate model output. Shown is an integrated water vapour (TMQ/prw) in $kg\ m^{-2}$.

What is an Atmospheric River?

Atmospheric River (AR) is a long and narrow structure of water vapour in the lower troposphere going outside of the tropics over a land mass[1].

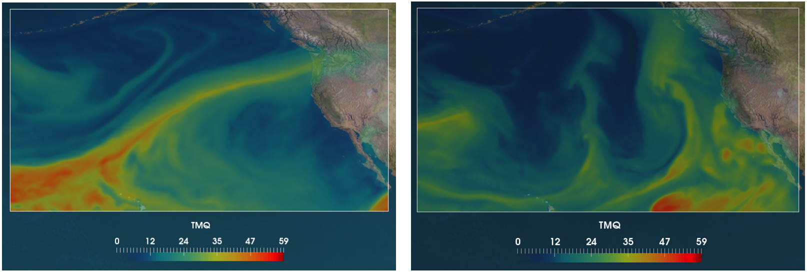


Figure 2: **Left:** An example of AR; **Right:** A snapshot of non-AR, i.e. having no filamentary structure. Shown is an integrated water vapour (TMQ/prw) in $kg\ m^{-2}$.

[1] Newell, Reginald E., et al. "Tropospheric rivers?–A pilot study." Geophysical research letters 19.24 (1992): 2401-2404.

Goals of the project:

- **avoiding selection of subjective thresholds** on physical variables in the detection scheme, like *TMQ* variable, *i.e.* Integrated Water Vapour (IWV) in $kg\ m^{-2}$.
- **providing reliable AR pattern detection method** that works for **different resolutions** of climate models.
- **identifying AR patterns** with high accuracy and precision.

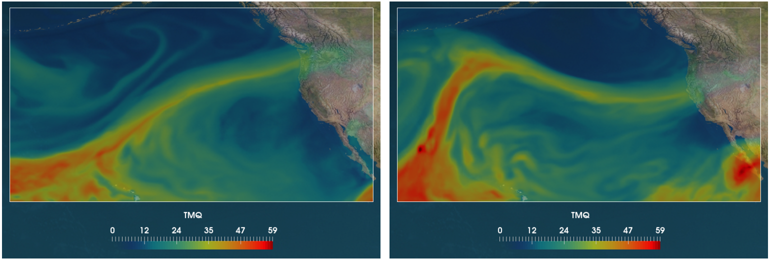


Figure 3: ARs can have very different shapes. Shown is an integrated water vapour (TMQ/prw) in $kg\ m^{-2}$.

AR pattern detection method

We can distinguish the two stages:

- Stage 1: feature extraction - Topological Data Analysis (TDA), i.e. **Union-Find algorithm**;
- Stage 2: classification task - Machine Learning, *i.e.* **Support Vector Machine classifier**.

Input and Output of the method:

- **Input:** scalar fields on a 2D regular grid.
- **Output:** a set of binary labels: AR=1, otherwise non-AR=0.

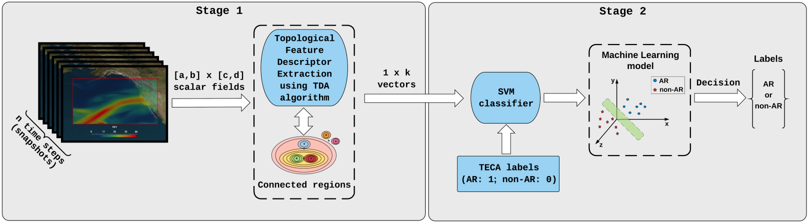


Figure 4: Block diagram of the AR patterns detection method.

Stage 1: Feature extraction

- Extracting numerical features of topological descriptors called *connected components (regions)*. The core of Union-Find algorithm is a disjoint set data structure[2].

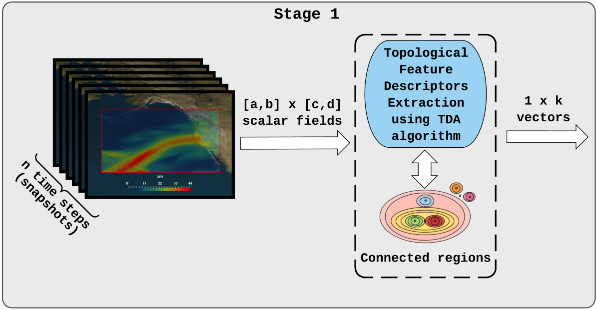


Figure 5: An illustration of Stage 1 of the method: extracting topological feature descriptors from 2D scalar fields on a grid.

[2] Hopcroft, John E., and Jeffrey D. Ullman. "Set merging algorithms." SIAM Journal on Computing 2.4 (1973): 294-303.

Stage 1: Climate data representation

Climate model output may be represented as a mapping from the grid to a set of real values and it can be defined as follows

$$f : [a, b] \times [c, d] \to [0, L], \tag{1}$$

where a, b, c and d are the dimensions of the grid and L is the maximal value of a variable (here IWV, $L = 60\ kg\ m^{-2}$).

Every grid point has four neighbours in the grid (except boundary points), i.e. the point at $(x, y) \in [a, b] \times [c, d]$ has four neighbours that have the coordinates $(x \pm 1, y)$ or $(x, y \pm 1)$.

This is the so-called 4-connected neighbourhood!

Stage 1: Union-Find algorithm

Following the threshold-free approach in TDA[3], the evolution of connected regions in a *superlevel set* is monitored at every value t of function f.

The superlevel set is a set of grid points in the domain of function f with scalar value greater than or equal to t. It is possible to mathematically express the superlevel set as follows

$$f^{-1}[t, +\infty) = \{(x, y) \in [a, b] \times [c, d] : f(x, y) \geq t\}. \qquad (2)$$

As t is decreased connected regions of $f^{-1}[t, +\infty)$ start to appear and grow and eventually merge into larger components.

The computational time complexity is $\mathcal{O}(n \log n)$, where n is the number of grid points.

---

[3] It is inspired by *persistence*, which is a concept in TDA that summarizes topological variations across all values of the scalar field under consideration.

Stage 1: a toy example

Suppose there are three connected regions (C_0, C_1, C_2) at value t_0 in a superlevel set.

As values of f decrease, the component C_0 grows until eventually, at t_1, it merges into the component of C_1, which in turn, merges into the component of C_2 at t_2, and so on.

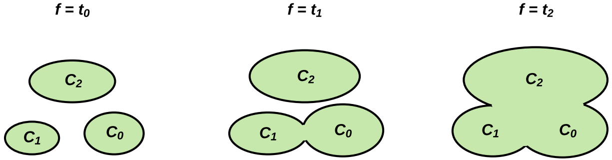


Figure 6: An illustration of the connected regions in the superlevel set.

Stage 1: a real data example

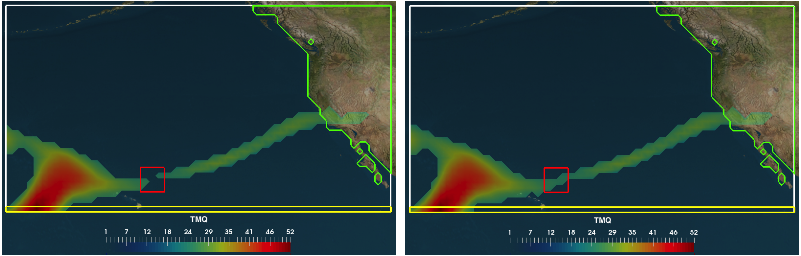
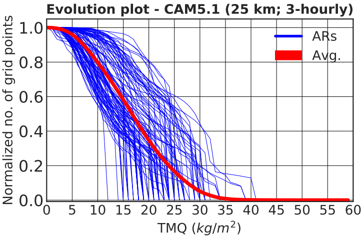


Figure 7: An illustration of finding connected AR regions over a specified search sector. In this example, the search for ARs is bounded by the latitude of the Hawaiian Islands (yellow line) and the west coast of North America (green line). **Left:** The red box indicates location of two regions that are disconnected at some value $IWV = t_1$. **Right:** At a new value $IWV = t_2$, where $t_2 < t_1$, the two connected regions merge into one new connected region forming a valid AR pattern.

Stage 1: an output of the algorithm applied to real data

Our algorithm monitors changes in superlevel sets (*i.e.*, special case of level set approach) connecting two geographical locations (*e.g.*, lat. of Hawaii and the western coast of US).



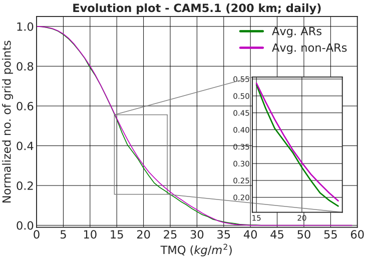


Figure 8: Topological feature descriptors of: **Left**: 100 randomly selected AR snapshots; **Right**: Averaged and normalized topological descriptors for all dataset.

Stage 1: creating an input for the Stage 2

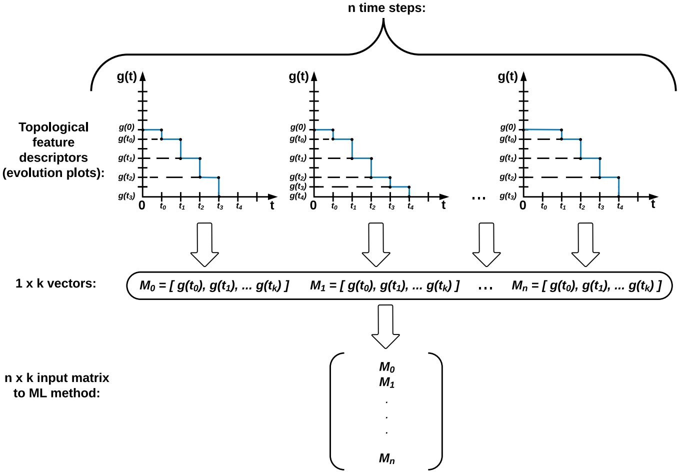


Figure 9: Mapping evolution plots, where $g(t)$ is the number of grid points the connected component for each threshold value t of TMQ/prw, n is the number of snaphots and k is the maximal value of TMQ/prw variable in a given climate model output.

Preprocessing of data

- Data normalization (standardization) is a way of adjusting measured values to a common scale (i.e., $[0, 1]$) by dividing through the largest maximum value in each feature (column of the matrix).
- Balancing data is motivated by the imbalanced class problem, which is that each class of event (ARs and non-ARs) is not equally represented in the dataset. Resampling has been applied to the output produced by the Union-Find algorithm along with labels provided by TECA, i.e. Toolkit for Extreme Climate Analysis.

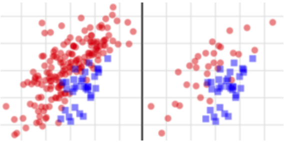


Figure 10: Balancing data is achieved by applying the resampling technique called undersampling: **Left**: Original dataset. **Right**: Resampled dataset.

Stage 2: AR detection by Support Vector Machine (SVM)

Detection of ARs is formulated as a binary classification task that requires the following steps:

- Incorporating labels for training process of the SVM classifier.
- Using exhaustive hyper-parameters grid searching, *i.e.* loose and fine grid searching approaches are applied.
- Performing cross-validation classification.

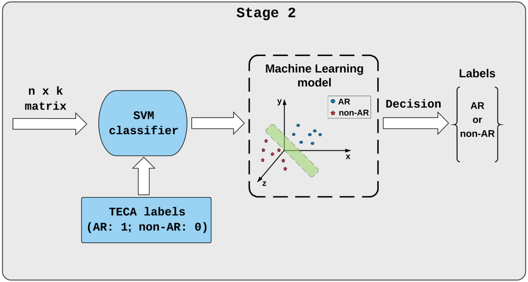


Figure 11: Classifying ARs based on topological feature descriptors extracted from 2D scalar fields and labels provided by TECA.

Stage 2 cont'd. - A separable two-class dataset

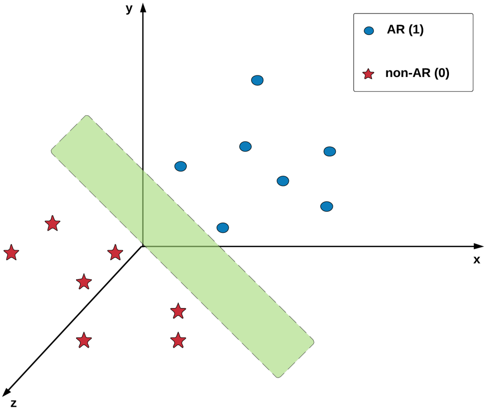


Figure 12: An example of a two-class dataset that is separable in some high-dimensional feature space.

Results: evaluation metrics

- Confusion Matrix:

	Label non-AR	**Label AR**
Predicted non-AR	True negatives	False positives
Predicted AR	False negatives	True positives

Figure 13: A confusion matrix (error matrix) is a way to present the performance of the method (typically, testing accuracy).

- Classification Accuracy Score: $\frac{Tp+Tn}{Tp+Tn+Fp+Fn}$
- Precision Score: $\frac{Tp}{Tp+Fp}$
- Sensitivity Score: $\frac{Tp}{Tp+Fn}$

Results: classification accuracy

The obtained accuracy is up to **91%**.

Dataset	Training Accuracy	Testing Accuracy	# of AR snapshots	# of Non-AR snapshots
CAM5.1 (25 km)	83%	83%	6838	6848
CAM5.1 (100 km)	77%	77%	7182	7581
CAM5.1 (200 km)	90%	90%	3914	3914

Dataset	Training Accuracy	Testing Accuracy	# of AR snapshots	# of Non-AR snapshots
CAM5.1 (25 km)	78%	82%	624	624
CAM5.1 (100 km)	85%	84%	700	700
CAM5.1 (200 km)	89%	91%	397	397

Dataset	Training Accuracy	Testing Accuracy	# of AR snapshots	# of Non-AR snapshots
MERRA-2 (50 km)	80%	80%	13294	13434

Figure 14: List of datasets used to test the method.

Results: precision and sensitivity

The obtained precision is up to **0.97**.

Dataset	Precision	Sensitivity
CAM5.1 (25km, 3-hourly)	0.91	0.74
CAM5.1 (100km, 3-hourly)	0.83	0.67
CAM5.1 (200km, 3-hourly)	0.95	0.85
CAM5.1 (25km, daily)	0.87	0.77
CAM5.1 (100km, daily)	0.86	0.83
CAM5.1 (200km, daily)	0.97	0.85
MERRA-2 (25km, 3-hourly)	0.84	0.74

Figure 15: Precision and sensitivity scores for all datasets.

Limitations of the method

- The method might fail if there are present two separate events (see left panel in Figure 24).
- The method might fail due to imperfect training data (see right panel in Figure 24).

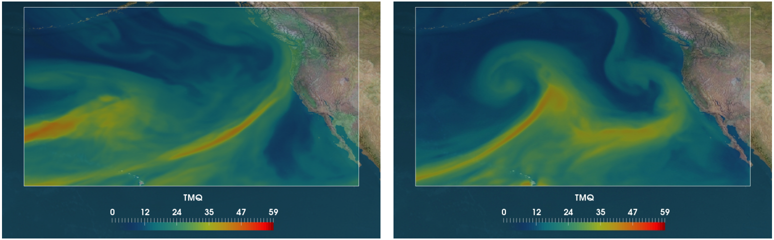


Figure 16: Sample images of events from the testing set showing a typical failure mode of the proposed method.

Conclusions

- **We have proposed a new way of analyzing weather patterns**, in particular Atmospheric Rivers.
- The presented method is **threshold-free** and **adaptable to different climate model's resolutions**.
- **The Union-Find algorithm** reduces the feature extraction process to **couple of minutes** in comparison with training of Convolutional Neural Networks (*i.e.*, days or weeks);
- The proposed method has achieved a high classification accuracy and precision up to **91%** and **0.97**, respectively.

Future Work

- We consider applying the method to direct observations, *i.e.* Special Sensor Microwave Imager/Sounder (SSMIS) satellite images.
- We plan to design a characterization and detection framework for Atmospheric Blocks.
- The framework is based on Manifold Learning and Topological Data Analysis/Machine Learning.

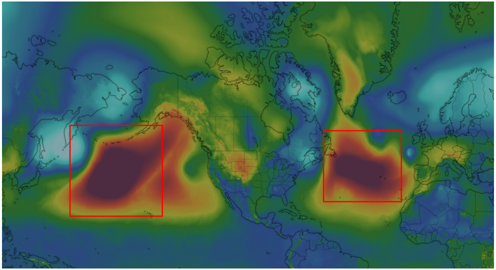


Figure 17: An example image of two Atmospheric Blocks.

Acknowledgments

Thanks to:

- Dmitriy Morozov (LBNL)
- Burlen Loring (LBNL)
- Hari Krishnan (LBNL)
- *Intel* for supporting this project through the Big Data Center at the Berkeley Lab, US.

References

- Shields, C. A., et al.: **Atmospheric River Tracking Method Intercomparison Project (ARTMIP): Project Goals and Experimental Design**, Geosci. Model Dev.
- Muszynski, G., Kashinath, K., Kurlin, V., Wehner, M., Prabhat: **Topological Data Analysis and Machine Learning for Recognizing Atmospheric River Patterns in Large Climate Datasets**, Geosci. Model Dev., **(in review)**, https://doi.org/10.5194/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, **(in review)**.