Topological Methods for Pattern Detection in Climate Data

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

[1]Lawrence Berkeley National Laboratory, Berkeley, California, 94720, United States
[2]Department of Computer Science, University of Liverpool, L69 3BX, United Kingdom

Key Points:

- A method for identifying extreme weather patterns in large climate simulation products is described.
- The method is free of manually chosen threshold parameters on physical variables.
- This method is based on applied topology and machine learning framework.

Corresponding author: Grzegorz Muszynski, muszynski.grzegorz.tda@gmail.com

Abstract

Nowadays, massive climate simulation datasets are produced due to the unprecedented increase in computing power, and there is a need to provide automated methods for analyzing these data. Here we focus on one particular class of methods, *i.e.* methods for local detection of extreme weather phenomena. We describe an automated method for the identification of the extreme events in large sets of climate simulation data. This method adapts an algorithm for topological data analysis to extract numerical features of topological descriptors called connected components. The features are then fed to a supervised machine learning classifier. The classifier performs a binary classification task to identify the extreme weather patterns we are interested in. We illustrate capabilities of this method by presenting a case study of atmospheric river patterns that are often associated with severe precipitations in the mid-latitudes. We also show that the method can be suitable for analyzing large amounts of climate simulation products. Hence, we think that climate community will find this example instructive and inspiring. We also indicate other future climate science problems in which applied topology coupled with machine learning can be found useful.

1 Introduction

Very complex climate models have been used to simulate physical processes of the global climate system. The primary goal of building climate models is to obtain numerical data that can advance our understanding of a changing climate. The higher level of detail in a climate model is, the more accurate the model becomes. For example, a detailed climate model can accurately capture the physical features of weather events. For this reason, the climate science community repeatedly runs simulations with different scenarios and resolutions using powerful supercomputers. This is possible due to the development of high performance computing infrastructure in the past decade. However, we now have great opportunities to run models with the higher spatial and temporal resolutions. As a result, data have become more complex, and our ability to analyze the models output has been outpaced by our ability to produce massive amounts of data. That is why there are many attempts to provide automated methods that can perform a rapid analysis of big climate data [*Lyubchich et al.*, 2017].

One of the challenges in climate data analysis is designing robust methods for the detection of extreme weather events, such as extra-tropical cyclones and atmospheric rivers [*Newell et al.*, 1992; *Shields et al.*, 2018] (see Figure 1). Identifying such weather patterns is important for: i) assessing climate models and ii) producing event statistics, *i.e.* the frequency, location and intensity of the events under global warming.

Recently, numerous automated methods for extreme event detection have been proposed by the climate and physics community. Nonetheless, most of the methods are based on arbitrary threshold parameters and do not work as well as human domain experts. Some of the newest machine learning methods, such as deep learning techniques which circumvent choosing critical threshold conditions, have been used for the detection of extreme weather patterns [*Racah et al.*, 2017]. However, the training process for deep learning methods to capture features of data is not sufficiently understood and is time consuming. Hence, there is much of ongoing research to provide fast and still arbitrary threshold-free methods for the extreme weather event detection.

In this chapter, we describe an alternative approach to extreme weather pattern detection. In particular, we present an automated method that is based on an applied topology algorithm and a supervised machine learning classifier. The method adapts the recent advances in topological data analysis (TDA) that is an emerging branch of data science [*Munch*, 2017; *Patania et al.*, 2017]. TDA provides feature extraction algorithms using techniques of topology and computer science to study intrinsic properties of data [*Carlsson*, 2014]. Here we focus on a particular algorithm called Union-Find (U-F) that provides a unique and threshold-free way of describing the crucial shape characteristics of physical phenomena. This algorithm computes numerical features of topological descriptors, *i.e.* connected components in 2D scalar field data [*Edelsbrunner and Harer*, 2010]. In this study, the-U-F algorithm ex-

tracts numerical features of the descriptors of a given scalar field on a latitude-longitude grid. The extracted features from positive (the extreme events) and negative (those that are not the extremes) examples are then used in the training process of a machine learning classifier, *i.e.* Support Vector Machine (SVM) [*Chang and Lin*, 2011]. The trained classifier performs the task of binary classification for recognizing weather patterns in climate simulation output. Furthermore, we show an application example of the method in a case study of atmospheric river events (ARs) in the Community Atmosphere Model (CAM5) output [*Muszynski et al.*, 2019]. ARs are very often associated with heavy precipitation in the mid-latitudes (*e.g.* the western coast of United States) and those making landfall are less frequent than any other events. Figure 1 shows an example of AR that deposits large amounts of rainfall on California (see left image). The method is applied to ARs making landfall along the western coast of North America, but it can be easily extended to other regions.

The rest of the chapter is organized as follows: Section 2 describes the topological based U-F algorithm and a supervised machine learning classification task, including the SVM classifier; Section 3 shows the results obtained in the case study of ARs and Section 4 presents conclusions and recommendations for readers.

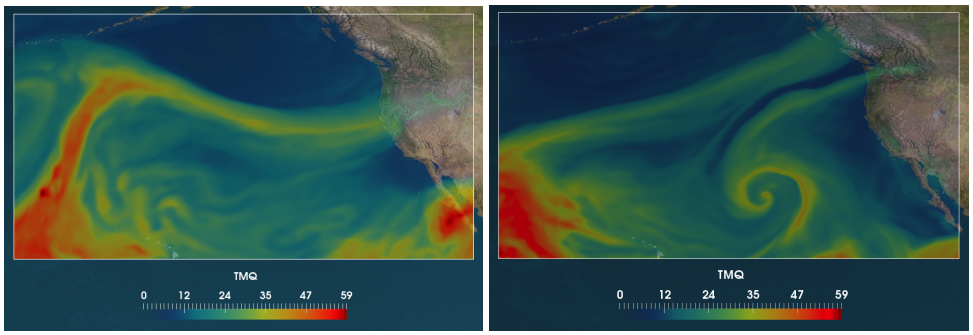

Figure 1: Sample images of two weather patterns having distinguishable structure in climate model output. **Left**: Atmospheric River (a long filamentary structure stretching from Hawaii Islands to the western coast of United States) making landfall over California State, United States. **Right**: Extra-Tropical Cyclone (a spiral shape structure) approaching the western coast of United States. Shown is total precipitable water (TMQ in $kg\ m^{-2}$) from a simulation of 5.1 version of the Community Atmosphere Model (CAM5.1).

2 Topological Methods for Pattern Detection

The main goal of this section is to explain the topological and machine learning based method for extreme weather pattern detection in climate simulation output. Firstly, the U-F algorithm is described [*Hopcroft and Ullman*, 1973]. The algorithm is the foundation on which the method is built. Next, we will introduce a supervised machine learning approach with emphasis on the SVM classifier that is commonly used in a binary classification tasks [*Chang and Lin*, 2011]. To sum up, the method consists of two steps:

- **Step 1**: The U-F algorithm automatically extracts numerical features of topological descriptors called connected components. The features of the descriptors are obtained from 2D scalar fields (snapshots) of global climate model output on a latitude-longitude grid. The numerical features provided by the algorithm are then used as the input matrix for a machine learning classifier in the Step 2.
- **Step 2**: This step employs the machine learning classifier (SVM) to perform detection which is formulated as a binary classification task. There are two stages in the classification task: 1) Training the classifier on the numerical features (from Step 1) with la-

bels. The classifier learns how to distinguish extreme events of our interest from other weather events in the snapshots. 2) Testing the trained classifier on the unlabeled numerical features ("held out data"). The classifier separates events into two groups, *i.e.* class A: the extreme weather patterns (1) and class B: those that are not (0).

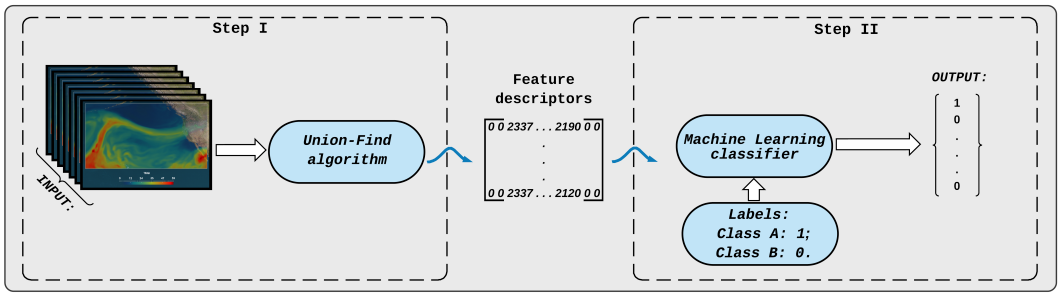

Figure 2: The block diagram illustrating the extreme weather pattern detection method. The input of the method is a set of scalar fields (snapshots) on the latitude-longitude grid. The Union-Find algorithm extracts numerical features of connected components from the snapshots of global images on the grid. The obtained matrix feeds into machine learning classifier with labels, *i.e.* Class A: the extreme weather patterns (1) and Class B: those that are not (0). The classifier is taught how to cleanly separate events into two groups. Finally, the output of the method is a set of labels based on the decision made by the classifier on unlabeled data.

2.1 Step 1: Topological Feature Descriptors of Weather Patterns

The goal of this step is to automatically produce numerical features of topological descriptors from 2D snapshots (scalar fields) in climate model output. Most of existing methods for extreme weather pattern detection rely on choosing subjective thresholds. The approach proposed here is inspired by TDA, in particular "persistence". It is a concept in applied topology that summarizes topological variations across all values of the scalar field under consideration (free of threshold conditions) [*Ghrist*, 2008].

Climate model output is usually a mapping (function) from the grid to a set of real values. In this case it is a variable over $[0, V]$, where V is the maximal value of the variable. Formally, it can be defined as function $f : [a, b] \times [c, d] \rightarrow [0, V]$, where a, b, c and d are the dimensions of the grid. Every point in the grid $(x, y) \in [a, b] \times [c, d]$ has four neighbouring points (except those lying on the boundaries). Each neighbour can have the coordinates $(x \pm 1, y)$ or $(x, y \pm 1)$, *i.e.* the 4-connected neighbourhood.

The evolution of connected components in a superlevel set $f^{-1}[t, +\infty) = \{(x, y) \in [a, b] \times [c, d] : f(x, y) \geq t\}$ is monitored at every value t of the function f. As t decreases, the components in the superlevel set $f^{-1}[t, +\infty)$ start to appear and grow, and eventually merge into one component covering the entire domain of the function f. This is the so-called threshold-free approach in the TDA.

Here is an illustration of this approach. Suppose that there are three connected components C_0, C_1 and C_2 at t_0 in the superlevel set $f^{-1}[t, +\infty)$, as is shown in Figure 3. As values of f decrease, the component C_0 grows until eventually merges into the component of C_1 at t_1, after which, it merges into the component of C_2 at value t_2.

The discussed threshold-free approach of connected components can be performed by the U-F algorithm based on disjoint set data structure [*Hopcroft and Ullman*, 1973]. The algorithm finds connected components of a grid by operating on sorted grid points by scalar values in decreasing order. The disjoint set data structure maintains the components and keeps track of the evolution of these components in the grid.

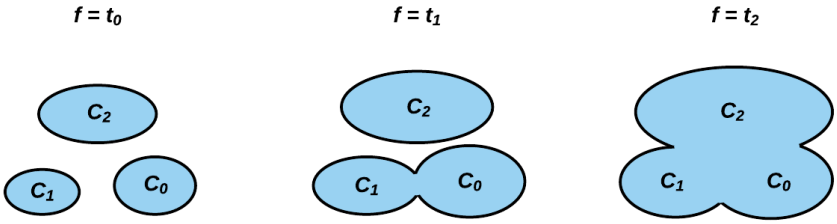

Figure 3: A toy example of three connected components (C_0, C_1, C_2) in the superlevel set, *i.e.* the three separate pieces at value t_0. The components grow and merge first at value t_1 and then at t_2 when values of function f are gradually decreased [*Muszynski et al.*, 2019].

When the U-F algorithm is applied to scalar fields on the latitude-longitude grid, there are five main operations used in the algorithm:

1. create a new connected components and add it to the disjoint set data structure;
2. assign a given grid-point to the right connected component;
3. check if the component intersects a specified geographical location on the grid;
4. merge two connected components containing at least one neighbouring grid point into one new connected component;
5. track the evolution of connected components intersecting a specified geographic location (number of grid points in it) as values of scalar field are systematically varied;

The extracted numerical features of connected components are encoded into evolution plots, as shown in Figure 4. The plots show the recorded number of grid points in the component as values of variable describing the scalar field are systematically decreased. The horizontal axis t contains values of variable and the vertical axis $g(t)$ shows number of grid points in the connected component. The information from the plots of each snapshot is encoded as a matrix (set of row vectors stacked on top of each other). This matrix is an input to the machine learning classifier, described in the next section.

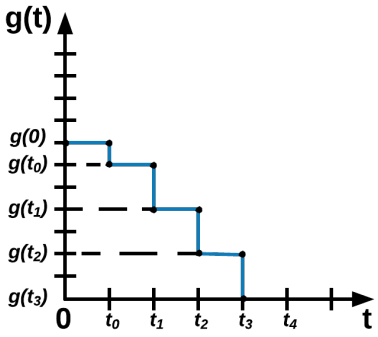

Figure 4: An example of evolution plot describing the changes of the connected components in the superlevel sets as variable t is systematically decreased [*Muszynski et al.*, 2019].

2.2 Step 2: Machine Learning for Classifying Weather Patterns

Machine learning approaches can be divided into three major categories: supervised, semi-supervised and unsupervised learning [*Kubat*, 2015]. Here we focus on the first one that incorporates labeled data in the process of training the machine learning classifier. The most common supervised learning tasks are regression and classification, with models including logistic regression, support vector machine, and deep learning models in use recently.

Here we focus on the support vector machine (SVM) as it is a widely used machine learning classifier for a binary classification task [*Chang and Lin*, 2011]. The main objective of SVM is to decide whether a particular weather pattern is present or not in a given snapshot extracted from global climate model output. The SVM constructs a model based on the labeled numerical features of topological descriptors in the training set. Next the model predicts the labels of the testing set consisting of the unlabeled descriptors. In general, the SVM finds the optimal hyperplane that separates two groups of patterns (Class A: 1 and Class B: 0) by maximizing the margin between the separating boundary and the training points closest to the support vector.

Assume a training set of instance-labels pairs (x_i, y_i), $i = 1, ..., N$, where $x_i \in \mathbb{R}^n$ and $y_i \in \{1, 0\}$. The solution of the optimization problem (finding the optimal hyperplane) is given by $\min_{w,b,\xi} (\frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i)$, subject to $y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$. The penalty parameter of the error term takes only values greater than zero ($C > 0$) and $\xi_i \geq 0$ is a minimum error when two groups are not linearly separable (*e.g.*, due to noise in training data). The samples x_i from training set are mapped into a higher dimensional space by the kernel function to make the samples of two groups (Class A and Class B) separable, as shown in Figure 5.

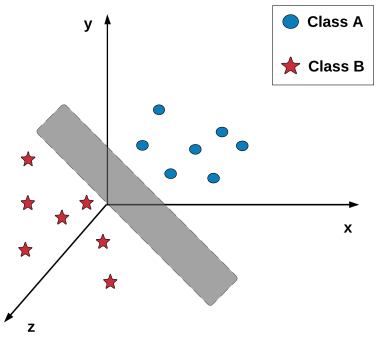

Figure 5: An illustration of a two-class data that is separable in a high-dimensional space. The input data set has been transformed into a high-dimensional feature space, such that in this space there exists a optimal hyperplane (gray surface in the figure) that cleanly separates the data into two groups, positive (Class A) and negative (Class B).

3 Case Study: Atmospheric Rivers detection

This section presents one of possible applications of the method to pattern detection of an extreme weather phenomenon called atmospheric river (AR). The method has been tested on output of version 5.1 of the Community Atmosphere Model (CAM5.1). The summary of the data is listed in the Section 3.2. Firstly, we compare the obtained numerical features of the topological descriptors based on the labels provided by the Toolkit for Extreme Climate Analysis (TECA) [*Prabhat et al.*, 2015]. Secondly, we estimate performance and reliability

of the method in the context of classification accuracy, precision and sensitivity score obtained by the SVM classifier.

3.1 Atmospheric Rivers

ARs are long filamentary atmospheric structures of high concentrated water vapour in the troposphere (see Figure 6). The climate science community often connects them with extreme precipitations in mid-latitudes [*Newell et al.*, 1992]. ARs are present on the western coast of North America and as well as along the Atlantic European coasts. They can pose a high risk to society by causing floods when they make landfall. The AMS glossary defines

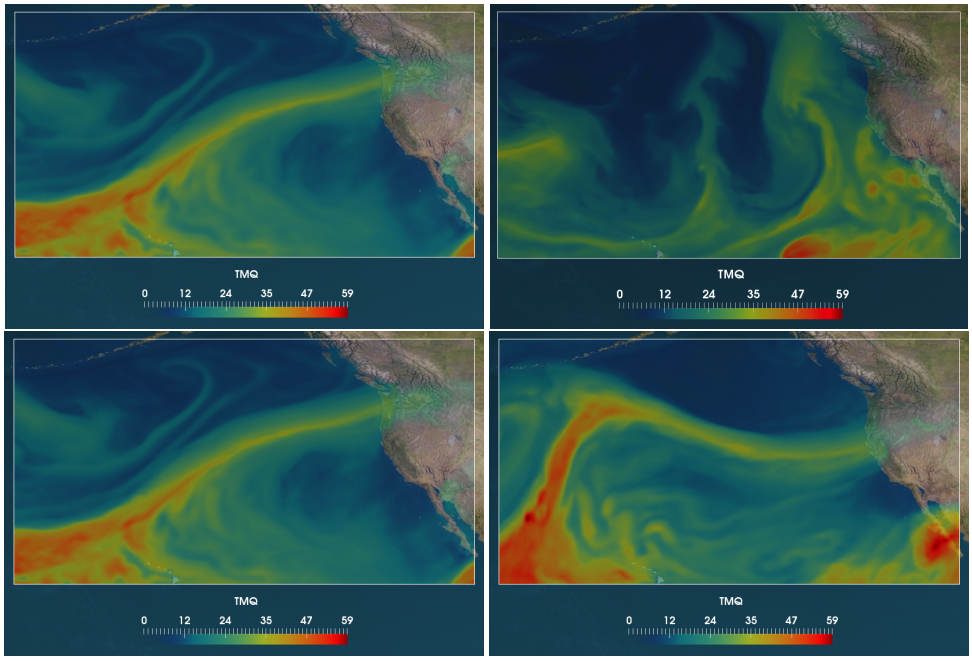

Figure 6: Sample images illustrating AR detection problem. The upper row shows an AR (left) and non-AR (right). The lower row shows two ARs having different geometric structure. Shown is integrated water vapor (TMQ in $kg\ m^{-2}$) from a simulation using the 5.1 version of the Community Atmosphere Model (CAM5.1).

an AR as follows: *A long, narrow, and transient corridor of strong horizontal water vapor transport that is typically associated with a low-level jet stream ahead of the cold front of an extratropical cyclone. The water vapor in atmospheric rivers is supplied by tropical and/or extratropical moisture sources. Atmospheric rivers frequently lead to heavy precipitation where they are forced upward, for example, by mountains or by ascent in the warm conveyor belt. Horizontal water vapor transport in the midlatitudes occurs primarily in atmospheric rivers and is focused in the lower troposphere* [*AMS*, 2018]. This definition is qualitative and numerous methods have been proposed to use it to detect ARs in regional and global climate data. But none of these methods are free from threshold conditions on a particular physical variable. Most of existing methods are based on a fixed threshold of more than 20 $kg\ m^{-2}$ of Integrated Water Vapour (IWV[1]) in the atmospheric column or more than 750 $kg\ m^{-2}$ of Integrated Water Vapour Transport (IVT). That is why choosing appropriate thresholds of IWV or IVT remains an open challenge [*Shields et al.*, 2018].

[1] For the CAM5.1 this variable is called *TMQ*. It is also called *prw* in the CF protocols.

3.2 Data

For the experiments, climate model simulation output generated by version 5.1 of the Community Atmosphere Model (CAM5.1[2]) has been used. The CAM5.1 climate model output is available at 25 km, 100 km, and 200 km spatial resolutions, and both 3-hourly and daily temporal resolutions, for the period of January 1979 to December 2005. Table 1 lists a summary of the model output. Both 3-hourly and daily data are used because the daily averages blur certain physical features of ARs. Moreover, 3-hourly output provides more event snapshots labeled as ARs, which is useful for training in the machine learning model[3].

3.3 Results

The extracted numerical features of topological descriptors (connected components) provide a unique way to characterize weather patterns (ARs) in climate model output (see Figure 7). The right and left plots correspond to ARs and non-ARs based on the provided TECA labels, respectively. Each curve represents the number of grid points in the connected component (superlevel set connecting two geographic locations) measured by the Union-Find algorithm. It other words, the algorithm records the evolution of the connected components as a function of the scalar variable (TMQ). We observe that is difficult to distinguish differences between these sets of curves. However, it is possible to train a machine learning classifier (SVM) to differentiate ARs and non-ARs with high accuracy.

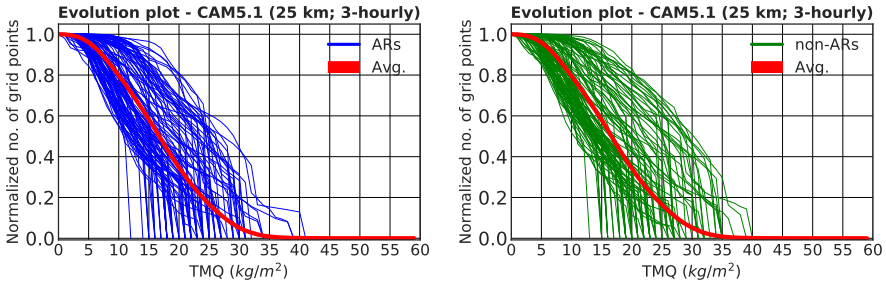

Figure 7: Examples of normalized evolution plots of averaged (bold curves) and 100 arbitrarily selected topological feature descriptors of ARs (right plot) and non-ARs (left plot). For 3-hourly temporal resolution and 25 km spatial resolution of the CAM5.1 simulation data. The plots illustrate how numerical features of topological descriptors (number of grid points in the connected component) vary with changing *TMQ* or *IWV* values [*Muszynski et al.*, 2019].

Tables 2 and 3 summarize the classification accuracy of the SVM. Training accuracy measures how well the model learns from training data (data labeled with ARs and non-ARs). Testing accuracy measures how well the method performs on a unlabeled dataset.

Table 2 shows that the SVM classifier is able to best differentiate ARs from non-ARs when the spatial resolution of the 3 hourly and daily climate model is low. Despite the fact that the high-resolution version of the model more accurately represents AR statistics, the IWV fields tend to be noisy, leading to a less smooth topological representation and lower training accuracy. Although, even with a low number of ARs available to train the SVM, the high testing classification accuracy for the CAM5.1 (200 km) suggests that the classifier is able to capture significant nonlinear dependencies between the features of topological descriptors.

[2] CAM5.1 data are provided by National Energy Research Scientific Computing Center (NERSC) at the Lawrence Berkeley National Laboratory (LBNL).

[3] Machine learning models achieve better results when more labeled data are available.

Table 3 shows the precision and sensitivity scores. Both scores can reach its best value at one and worst value at zero. They measure the method's ability not to classify as AR an event that is non-AR, and how well this method is in finding all the AR events. The method has the highest precision and sensitivity scores for 200 km resolution of CAM5.1 model for both 3-hourly and daily temporal resolutions. The scores are slightly lower for other spatial and temporal resolutions of CAM5.1 model. This suggests that the SVM using the topological features is reliable method to classify ARs and non-ARs.

4 Conclusions and Recommendations

In this chapter, we present one possible application of topological methods coupled with machine learning for identifying weather patterns. In particular, we show the use of the method to recognize atmospheric river events in big climate data.

We demonstrate that the method is reliable and achieves high accuracy when tested on a wide range of CAM5.1 climate model resolutions. We also observe that this method performs better for low-spatial-resolution simulation datasets than high-spatial-resolution datasets. Because the high-resolution datasets contain usually noisier AR patterns, for example, the presence of other events, like an extra-tropical cyclone. In this situation, the SVM model can be confused, and it likely fails.

The main advantage of incorporating the topological algorithm in this work is that it allows for a threshold-free analysis, which is not possible with most existing ARs detection methods. Furthermore, the presented method is much faster than using (e.g., convolutional neural networks [*Liu et al.*, 2016]). The processing time for the method is minutes versus a few days for the neural networks.

Topological methods are not only applicable to a 2-dimensional scalar field on a regular grid. It is possible to apply them to higher-dimensional or multivariate fields. That is why we anticipate the applied topology and machine learning framework could be an effective way to characterize and identify a wide range of other weather patterns, such as tropical cyclones or blocking events.

Acknowledgments

Grzegorz Muszynski and Vitaliy Kurlin would like to acknowledge Intel for supporting the Intel Parallel Computing Center (IPCC) at University of Liverpool. Karthik Kashinath was supported by the Intel Big Data Center, and Michael Wehner was supported by the Regional and Global Climate Modeling Program of the Office of Biological and Environmental Research in the Department of Energy Office of Science under contract number DE-AC02-05CH11231. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

AMS (2018), Atmospheric River. Glossary of Meteorology, available online at `http://glossary.ametsoc.org/wiki/Atmospheric_river`.

Carlsson, G. (2014), Topological pattern recognition for point cloud data, *Acta Numerica*, *23*, 289–368.

Chang, C.-C., and C.-J. Lin (2011), Libsvm: a library for support vector machines, *ACM transactions on intelligent systems and technology (TIST)*, *2*(3), 27.

Edelsbrunner, H., and J. Harer (2010), *Computational topology: an introduction*, American Mathematical Soc.

Ghrist, R. (2008), Barcodes: the persistent topology of data, *Bulletin of the American Mathematical Society*, *45*(1), 61–75.

Hopcroft, J. E., and J. D. Ullman (1973), Set merging algorithms, *SIAM Journal on Computing*, *2*(4), 294–303.

Kubat, M. (2015), *An Introduction to Machine Learning*, Springer.

Liu, Y., E. Racah, J. Correa, A. Khosrowshahi, D. Lavers, K. Kunkel, M. Wehner, W. Collins, et al. (2016), Application of deep convolutional neural networks for detecting extreme weather in climate datasets, *arXiv preprint arXiv:1605.01156*.

Lyubchich, V., N. Oza, A. Rhines, E. Szekely, I. Ebert-Uphoff, C. Monteleoni, and D. Nychka (2017), Proceedings of the 7th international workshop on climate informatics: Ci 2017. ncar technical note ncar/tn-536+proc, pp. 240–241, doi:10.5065/D6222SH7.

Munch, E. (2017), A user's guide to topological data analysis, *Journal of Learning Analytics*, *4*(2), 47–61.

Muszynski, G., K. Kashinath, V. Kurlin, M. Wehner, and Prabhat (2019), Topological data analysis and machine learning for recognizing atmospheric river patterns in large climate datasets, *Geoscientific Model Development*, *12*, 613–628.

Newell, R. E., N. E. Newell, Y. Zhu, and C. Scott (1992), Tropospheric rivers?–a pilot study, *Geophysical Research Letters*, *19*(24), 2401–2404.

Patania, A., F. Vaccarino, and G. Petri (2017), Topological analysis of data, *EPJ Data Science*, *6*(1), 7.

Prabhat, S., Byna, V. Vishwanath, E. Dart, M. Wehner, W. D. Collins, et al. (2015), Teca: Petascale pattern recognition for climate science, in *International Conference on Computer Analysis of Images and Patterns*, pp. 426–436, Springer.

Racah, E., C. Beckham, T. Maharaj, S. Ebrahimi Kahou, M. Prabhat, and C. Pal (2017), Extremeweather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events, pp. 3405–3416.

Shields, C., J. Rutz, L.-Y. Leung, R. Martin, M. Wehner, et al. (2018), Atmospheric river tracking method intercomparison project (artmip): project goals and experimental design, *Geoscientific Model Development*, *11*(6), 2455–2474.

Table 1: List of data sources used in the experiments. The table shows output of historical runs of CAM5.1 model [*Muszynski et al.*, 2019].

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

Table 2: Classification accuracy score of the SVM classifier for 3-hourly and daily temporal resolution of CAM5.1 model with three different spatial resolutions. Table also shows number of snapshots (# of events for each category: ARs and non-ARs) [*Muszynski et al.*, 2019].

Spat. & Temp. Res.	**Training Acc.**	**Testing Acc.**	**# AR snapshots**	**# Non-AR snapshots**
(25 km, 3 h)	83%	83%	6838	6848
(100 km, 3 h)	77%	77%	7182	7581
(200 km, 3 h)	90%	90%	3914	3914
(25 km, a day)	78%	82%	624	624
(100 km, a day)	85%	84%	700	700
(200 km, a day)	89%	91%	397	397

Table 3: Precision and sensitivity scores calculated for all datasets listed in Table 1. The scores show the ability of the SVM in assigning correct labels to instances of testing set [*Muszynski et al.*, 2019].

Spat. & Temp. Res.	**Precision**	**Sensitivity**
(25km, 3 h)	0.91	0.74
(100km, 3 h)	0.83	0.67
(200km, 3 h)	0.95	0.85
(25km, a day)	0.87	0.77
(100km, a day)	0.86	0.83
(200km, a day)	0.97	0.85