# TECA: Petascale Pattern Recognition for Climate Science

Prabhat[1], Surendra Byna[1], Venkatram Vishwanath[2], Eli Dart[1], Michael Wehner[1], and William D. Collins[1]

[1] Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[2] Argonne National Laboratory, Argonne, IL, USA

**Abstract.** Climate Change is one of the most pressing challenges facing humanity in the 21st century. Climate simulations provide us with a unique opportunity to examine effects of anthropogenic emissions. High-resolution climate simulations produce "Big Data": contemporary climate archives are $\approx 5PB$ in size and we expect future archives to measure on the order of Exa-Bytes. In this work, we present the successful application of TECA (Toolkit for Extreme Climate Analysis) framework, for extracting extreme weather patterns such as Tropical Cyclones, Atmospheric Rivers and Extra-Tropical Cyclones from TB-sized simulation datasets. TECA has been run at full-scale on Cray XE6 and IBM BG/Q systems, and has reduced the runtime for pattern detection tasks from years to hours. TECA has been utilized to evaluate the performance of various computational models in reproducing the statistics of extreme weather events, and for characterizing the change in frequency of storm systems in the future.

**Keywords:** pattern detection, climate science, high performance computing, parallel I/O, data mining, petascale

## 1 Introduction

Climate simulations provide us with an unprecedented view of the state of earth's present, and potential future climate under global warming. State of the art climate codes, such as the Community Atmosphere Model (CAM v5) [2], when run in 25-km spatial resolution with 6-hour data dumps, produce over 100TB from a single 25-year integration period. The current CMIP-5 archive [3], consisting of international contributions from a number of climate modeling groups consists of over 5PB of data; this dataset was mined extensively for the IPCC AR5 report [5]. It is anticipated that CMIP-6 dataset [8] will cross the exabyte threshold with 25-km model runs being the norm. Faced with this massive deluge of multi-variate, spatio-temporal data, sophisticated and scalable "pattern detection" tools are critical for extracting meaningful scientific insights.

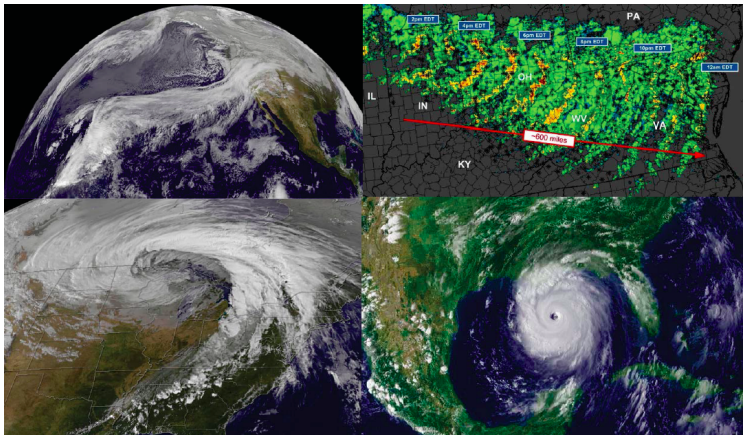


**Fig. 1.** Examples of extreme weather phenomena observed through satellite and radar. Clockwise from bottom-left: Extra-Tropical Cyclone, Atmospheric River, Derecho and Tropical Cyclone events.

One example of the types of climate data analytics of societal relevance is the identification and tracking of extreme weather. Figure 1 illustrates the types of extreme weather observed in the natural climate system. Phenomena such as cyclones and atmospheric rivers can have widespread and long-lasting impact on national economies. Understanding how extreme weather events will change in the future climate is an important open question.

In order to address this important challenge, we have developed the Toolkit for Extreme Climate Analysis (TECA)[12] to identify storms in high-frequency climate model output (CAM5 in our case). To date, we have applied our technique to identify three different classes of storms: tropical cyclones, atmospheric rivers and extra-tropical cyclones. Due to the high-frequency nature of the data required to identify and track individual storms in a climate model simulation, the raw input datasets that we have analyzed range from 0.5TB to 13TB. As the entire climate modeling community starts to upgrade their infrastructure to run at comparably high resolutions (25 km or better), we expect the publicly available datasets necessary for this type of analyses to exceed 10PB. Manual labeling and extraction of patterns at such scales is simply impossible, thereby requiring the development and application of "Big Data" analytics methods. The techniques that we have developed are amenable to parallel execution, and are demonstrated to scale up to full size of the largest machines available to us, including a 150,00 core Cray XE6 and 750,000 core IBM BG/Q platform.

## 2 Methods

### 2.1 TECA : Toolkit for Extreme Climate Analytics

TECA [12] is a climate-specific, high-performance pattern detection toolkit that is designed for efficient execution on HPC systems. We have developed the code in C/C++, and utilize MPI for inter-process communication. We utilize NetCDF-4 for parallel reads, and MPI-IO for storing results.

The design and implementation of TECA is based on our first hand experience with pattern detection problems in climate science. After analyzing the climate pattern detection literature for a number of event types, we discovered the following "design pattern". The detection process can be typically broken down into two steps:

1. Detection of candidate points that satisfy multi-variate constraints
2. Stitching of candidate points into a trajectory that satisfies spatio-temporal constraints

Step 1 tends to be data-intensive, involving loading anywhere between 10GB-10TB of data. The algorithm has to scan through all of the relevant fields to select candidate points. However, this step can be executed in parallel across timesteps. The degree of parallelism can be as high as the number of timesteps in the processed dataset (typically $10^2$-$10^5$); hence a dramatic speedup in overall runtime is feasible. Step 2 involves pairwise analysis on potential storm matches across consecutive time slices in order to stitch trajectories. However, a small amount of data (typically 10MB-1GB) is required for this analysis, and this can be easily loaded on memory on a single node and executed in serial.

Conceptually, Steps 1 and 2 can be directly translated to the MapReduce computational paradigm powering much of the commercial Big Data Analytics workloads. TECA implements a custom framework for processing scientific datasets, with an eye towards high performance. We utilize the Message Passing Interface (MPI) for optimizing job launch; communication and synchronization traffic. In the instance of multi-model archives, sub-communicators corresponding to individual models and ensemble members are created and initialized; thereby minimizing synchronization with other tasks. NetCDF files are striped across multiple low-level storage targets to optimize read performance. Writing (relatively small) partial results in Step 2 can create metadata bottlenecks, especially at concurrencies in excess of 50,000 cores. We implement a 2-phase collective I/O mechanism to aggregate writes on a smaller ($O(1000)$) number of nodes and perform file-per-node writes using MPI-IO. The cumulative effect of these best practices in HPC and Parallel I/O, is that we are able to successfully run TECA jobs at full concurrency on petascale platforms.

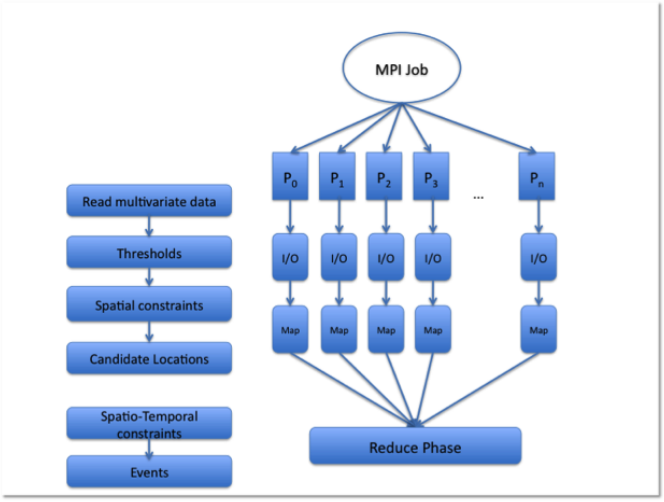


**Fig. 2.** TECA utilizes the Map Reduce computational paradigm for exploiting parallel computing resources.

### 2.2 TECA for detecting Tropical Cyclones

We have implemented the Tropical Cyclone (TC) detection procedure outlined in [11]. The detection step consists of finding co-located vorticity maxima, pressure minima (within a radius of 5°) and temperature warm-core centers. Splines are fitted to relevant fields, which are defined on latitude-longitude grids. Local maxima and minima of functions are computed by performing a line search on the splines within the prescribed spatial window. Once potential storm center candidates have been identified, the next step is to impose spatial and temporal constraints to connect the storm centers into a trajectory. In the case of TCs, the stitching step involves linking storms across subsequent 6-hr time windows. Candidate storms are prescribed to travel less than 400km in 6-hrs, persist for at least 2 days, and have a wind velocity (greater than 17m/s) during at least 2 days within their lifetime.

### 2.3 TECA for detecting Atmospheric Rivers

Atmospheric Rivers (AR) are large, spatially coherent weather systems with high concentrations of elevated water vapor. These systems often cause severe downpours and flooding over the western coastal United States and western Europe. We have implemented an algorithm to detect ARs in the TECA framework [9]. We first compute a 2D Integrated Water Vapor (IWV) by performing a vertical integral on the specific humidity field. Following the definition of physical

features of an AR [13]; we perform a thresholding operation for identifying all grid points with IWV greater than 2cm. We then use a connected component labeling algorithm to find all the connected regions of grid points. We test if a candidate originates in the tropics, and makes landfall on the US coast. For all the polygons satisfying the origin and the landfall conditions, we compute a medial axis, and check if the length of the AR greater than 2000km and if the width of the AR less than 1000km. If a polygon satisfies all of these geometric constraints, we declare it to be an atmospheric river.

### 2.4 TECA for detecting Extra-Tropical Cyclones

We implement the Extra-Tropical Cyclone (ETC) detection and tracking procedure documented in [14]. We detect a local minima in the pressure field within a 100x100 km radius. Ties between adjacent low-pressure storm centers are resolved based on strength of the local laplacian (i.e. storm centers with largest laplacian are declared to be storm center candidates). Potential candidates are stitched into trajectories by performing a nearest neighbor analysis with distance constraints: storms are restricted to travel less than 1000km in a 6-hr window, and less than 700 km in a 6-hr window in the North, South, and Westward directions. We only retain storms that persist for more than 24 hours, and travel greater 500km over their lifetime. Storms over high elevation areas (greater than 1500km) are excluded.

All of these detection and stitching criteria can be easily accommodated within the design of TECA; thereby utilizing parallel job launch, execution and parallel I/O capabilities. Apart from returning summary statistics on storm counts and location, we are also able to pull out valuable detailed information on precipitation patterns and velocity profiles of storms during the course of their lifetime.

## 3 Experimental Setup

### 3.1 Data

We utilized multi-model output from the community produced CMIP-5 archive [3], and a high-resolution version of the Community Atmospheric Model (CAM5)[2] simulations conducted by our group at NERSC. The CMIP-5 datasets are freely and publicly accessible via a number of international Earth System Grid Federation web portals and are the basis of most climate model results presented in the IPCC AR5 WG1 report [4]. The observational SSM/I datasets are available via a web portal [7]. In all cases, the datasets are available as multiple netcdf files. Each file typically contains all relevant 2D variables and spans one year's worth of data.

### 3.2 Platforms

We utilized the Hopper system at NERSC, and the Mira system at ALCF for all results reported in this paper. Hopper is a 1.28 PF, Cray XE6 system featuring

153,216 compute cores, 212TB of memory and 2PB of disk available via a 35 GB/s Lustre filesystem. Mira is a 10PF, IBM BG/Q system featuring 786,432 cores, 768 TB of memory with 384 I/O nodes accessible via GPFS.

## 4 Results

We now report on both the scaling performance obtained by TECA on various HPC platforms, as well as the scientific results facilitated by these runs.

### 4.1 Scaling Performance

Table 1 summarizes the performance of TECA on a range of pattern detection problems. We analyzed CAM5 model output (0.5-13 TB), CMIP-5 multi-model output (6 TB), and the SSMI (35 GB) satellite data product. We ran TECA at full scale on Hopper and Mira platforms, facilitating pattern detection on these massive datasets. Needless to say, such pattern detection problems cannot be tackled on individual workstations in a reasonable amount of time.

| Climate Pattern | Dataset | Dataset Size | Serial runtime (Estimated) | Parallel runtime | Concurrency | Platform |
|---|---|---|---|---|---|---|
| Tropical Cyclones | CAM5 1° | 0.5 TB | ≈ 8 years | 31 min | 149,680 cores | Hopper |
| Tropical Cyclones | CAM5 0.25° | 13 TB | ≈ 9 years | 58 min | 80,000 cores | Hopper |
| Atmospheric Rivers | SSM/I | 35 GB | ≈ 11 hours | 5 sec | 10,000 cores | Hopper |
| Extra-Tropical Cyclones | CMIP-5 | 6 TB | ≈ 10 years | 95 min | 755,200 cores | Mira |

**Table 1.** Scaling results obtained with TECA on various HPC platforms

### 4.2 Science Results

**Tropical Cyclones** One of the primary scientific utilities of the TECA software is to evaluate how well models perform in reproducing extreme event statistics, compared to observational records. If we assess models to perform well for the historical period, we can have greater confidence in the trends projected by the same models for future runs. We have applied TECA to the CAM5 0.25-degree output, over a simulated time period spanning 1979-2005 [15]. For this time period, the hand-labelled iBTrACS dataset [6] reports 87 (+/- 8) storms every year. TECA reports 84 (+/-9) storms, which is rather accurate. Figures 3 and 4 highlight the spatial distribution of the storms, as well as the seasonal distribution. In terms of model evaluation, we note that CAM5 does a good job

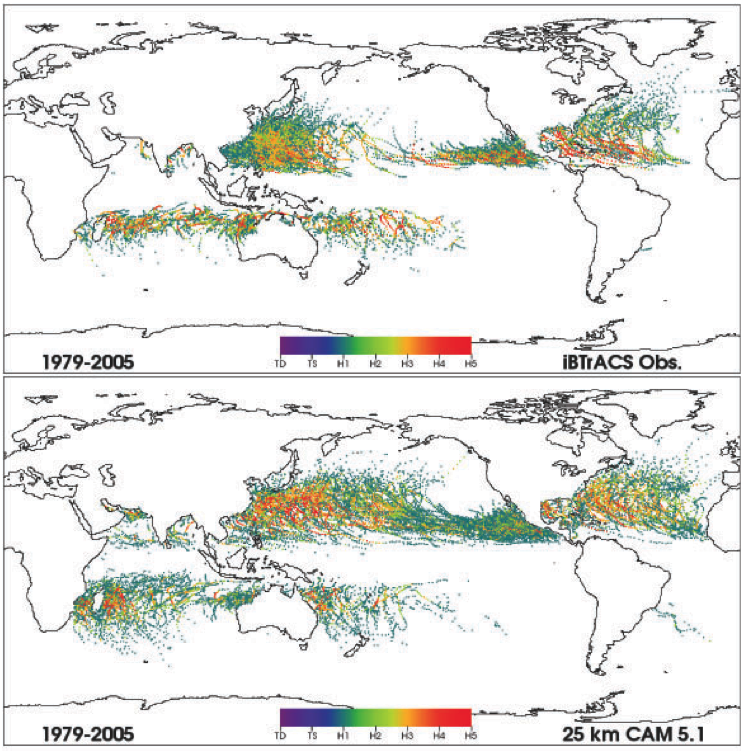


**Fig. 3.** Application of TECA to CAM5 0.25-degree output. Tropical Cyclones (Category 1 through 5) are illustrated in the bottom figure. TC tracks from the iBTraACS observational product are plotted for an identical time period.

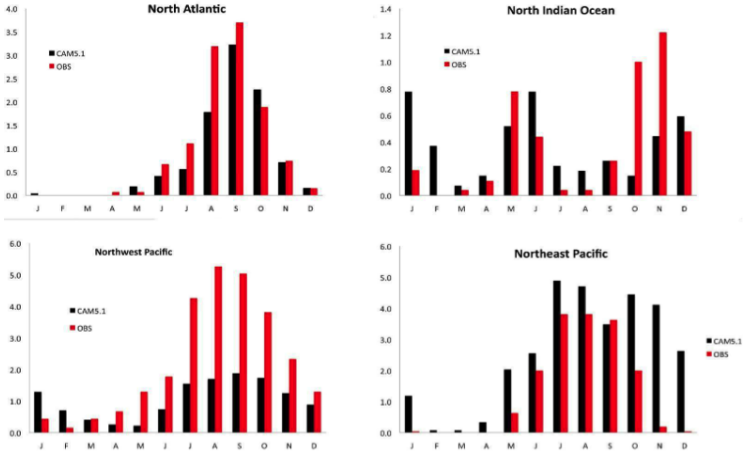


**Fig. 4.** TECA can produce detailed diagnostics for storm tracks. In this case, monthly TC activity is plotted by major oceanic basins. CAM5 output is plotted in black, and observational products are plotted in red.

of reproducing the spatial pattern, with perhaps too many storms in the central pacific. The model also does a good job of reproducing the seasonal pattern in various ocean basins (North Atlantic, Indian Ocean, Northwest Pacific), but the storms counts are off in the Pacific.

After validating the TECA output for the historical period, we decided to apply the TC detection capabilities for climate change experiments conducted by various US and international efforts. We processed a climate change experiment specified by the CliVAR Working Group [1]. We used the CAM5 model to simulated the earth's climate under a baseline (climo), a scenario consisting of 2xCO2, SSTs increased uniformly by 2°C, and the conjunction of both CO2 and SST conditions. Figure 5 shows the average number of tropical storms, tropical cyclones and intense tropical cyclones per year simulated by the high-resolution version (0.23°x0.31°) of CAM5.1 for the four idealized configurations. Error bars represent 5%-95% confidence intervals based on interannual variability. The baseline (1990) climatology is in blue. A two degree warmer simulation with elevated atmospheric carbon dioxide levels (660ppm) is shown in red. While the total number of tropical storms over all intensities is reduced in a warmer world, the number of intense tropical cyclones (Category 4 and 5) is increased.

**Atmospheric Rivers** We applied the TECA AR detection capability to the SSM/I satellite product. Figure 6 shows a range of diverse AR features returned

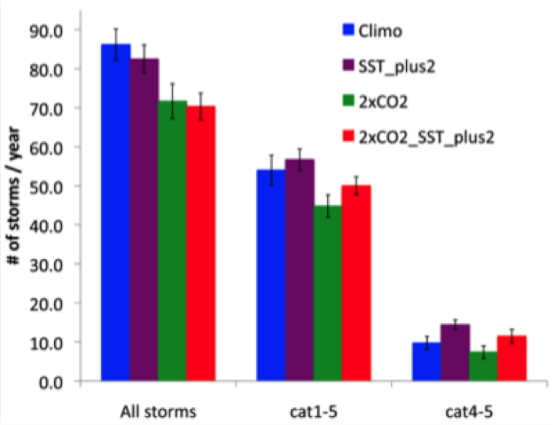


**Fig. 5.** Number of annual Tropical Cyclones under the CliVAR scenarios. There is a significant decrease in the overall number of storms, and an increase in number of annual Category 4 and 5 storms under the SST warming scenarios.

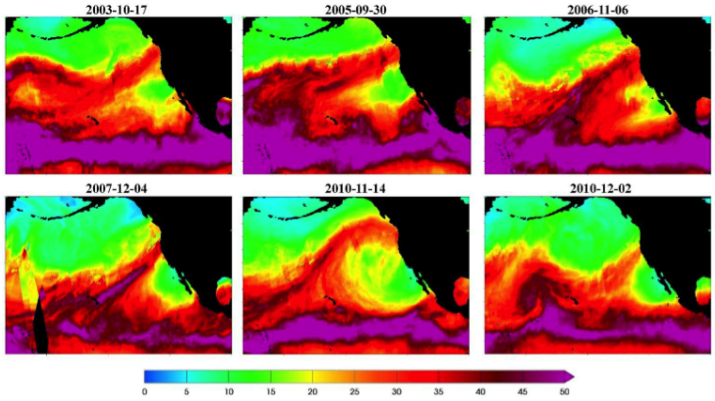


**Fig. 6.** Sample Atmospheric River events detected by the TECA implementation on the SSM/I dataset. Note the distinct appearance of various AR patterns detected by our algorithm.

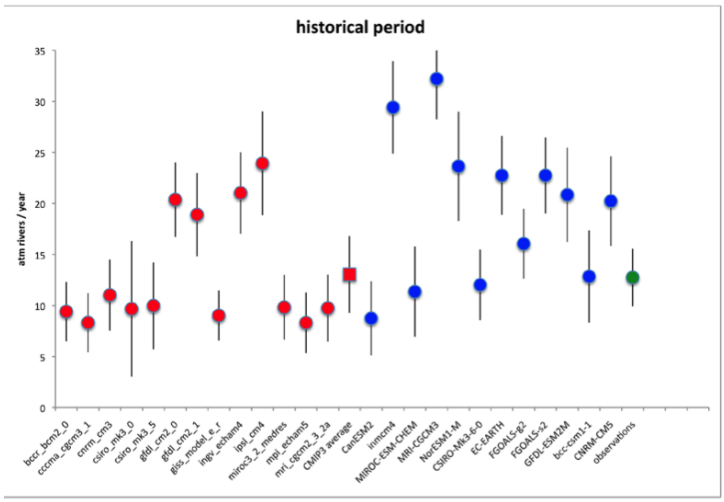


**Fig. 7.** Number of Atmospheric River events in the CMIP-3 (red) and CMIP-5 (blue) archives compared to observations (green). It appears that there are two modes in the CMIP-3 and CMIP5-archive: models that are generally consistent with the observed record, and models that are hyperactive in reproducing atmospheric rivers.

by our implementation. We note that the implementation is robust to various shapes and sizes of AR events. In order to validate the procedure, we compared the events returned by TECA to a hand-curated database of known AR events maintained by [10]. We note that TECA was able to detect 93% of all events reported in the database. We furthermore applied the TECA toolkit to various CMIP-3 and CMIP-5 models over the historical period. Figure 7 shows that several models match reasonably well with the observed record, however, some models do exhibit hyperactivity with regards to generation of ARs. Similar to the Tropical Cyclone work, we are currently investigating the application of TECA to climate change scenarios.

**Extra-Tropical Cyclones** We have successfully applied TECA to detect Extra-Tropical Cyclones in climate data. In perhaps the leading example of Scientific Big Data analytics, we scaled TECA to process the entire CMIP-5 archive (historical and RCP8.5 runs, all ensemble members, 6-hourly data) in one shot on 755,200 cores of the Mira IBM BG/Q system. Preliminary results in Figure 8 indicate that the extra-tropical cyclone count will decrease in a warming world, and that this trend is consistent across the entire CMIP-5 multi-model archive. We are currently designing custom aquaplanet simulations that will test specific hypothesis behind the decreasing trend in ETC activity.

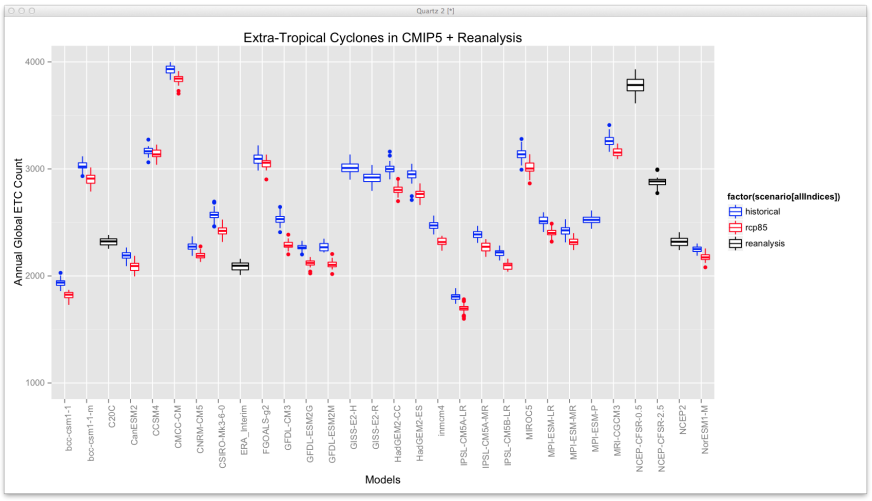


**Fig. 8.** Summary of Annual Extra-Tropical Cyclone activity in all of CMIP-5. A clear decrease is observed is observed from the blue (historical) to the future rcp8.5 (red) periods

## 5 Limitations and Future Work

This work is one of the leading example of the use of high performance computing for solving pattern "detection" problems. The definition of a pattern is assumed to be conveniently prescribed by various experts in the climate science community. While this is true for the three events that we examined, a much broader class of events exist (e.g. weather fronts, mesoscale convective systems, derechos, blocking events), for which there does not exist an algorithmic definition that could be implemented in TECA. Such problems are more amenable to the classic machine learning paradigm, wherein a modest number of training examples would be provided by human experts, and a learning algorithm could determine the relevant features and inter-relationships that "define" the extreme weather pattern. We are currently examining the use of Deep Learning methods, as a complementary approach to TECA for targeting a much broader class of weather patterns.

## 6 Conclusions

Pattern recognition problems are increasingly common in the scientific world. As a leading example, climate science requires sophisticated pattern recognition on TB-PB sized datasets. We have developed and successfully applied TECA to the problem of finding extreme weather phenomena (such as tropical cyclones, atmospheric rivers and extra-tropical cyclones) across most contemporary climate models (CAM5), data archives (CMIP-5) and observational products (SSMI).

We have scaled TECA on DOE's leading HPC platforms at NERSC and ALCF, and obtained important scientific insights on the potential change in extreme weather phenomena in future climate regimes.

## References


1. Clivar hurricane working group. http://www.usclivar.org/working-groups/hurricane.
2. Community earth system model. http://www.cesm.ucar.edu/working-groups/Atmosphere/development.
3. Coupled model intercomparison project phase 5. http://cmip-pcmdi.llnl.gov/cmip5/.
4. Earth system grid federation. http://pcmdi9.llnl.gov/esgf-web-fe/.
5. Intergovernmental panel on climate change, fifth assessment report. http://www.ipcc.ch/report/ar5.
6. Noaa ncdc ibtracs dataset. http://www.ncdc.noaa.gov/ibtracs/.
7. Remote sensing systems special sensor microwave imager instrument. http://www.remss.com/missions/ssmi.
8. Wcrp coupled model intercomparison project phase 6. http://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6.
9. S. Byna, Prabhat, M. F. Wehner, and K. J. Wu. Detecting atmospheric rivers in large climate datasets. In *Proceedings of the 2Nd International Workshop on Petascal Data Analytics: Challenges and Opportunities*, PDAC '11, pages 7–14, New York, NY, USA, 2011. ACM.
10. M. D. Dettinger, F. M. Ralph, T. Das, P. J. Neiman, and D. R. Cayan. Atmospheric rivers, floods and the water resources of california. *Water*, 3(2):445–478, 2011.
11. T. R. Knutson, J. J. Sirutis, S. T. Garner, I. M. Held, and R. E. Tuleya. Simulation of the recent multidecadal increase of atlantic hurricane activity using an 18-km-grid regional model. *Bulletin of the American Meteorological Society*, 88(10):1549–1565, 2007.
12. Prabhat, O. Rbel, S. Byna, K. Wu, F. Li, M. Wehner, and W. Bethel. Teca: A parallel toolkit for extreme climate analysis. *Procedia Computer Science*, 9(0):866 – 876, 2012. Proceedings of the International Conference on Computational Science, 2012.
13. F. M. Ralph, P. J. Neiman, and R. Rotunno. Dropsonde observations in low-level jets over the northeastern pacific ocean from caljet-1998 and pacjet-2001: Mean vertical-profile and atmospheric-river characteristics. *Monthly weather review*, 133(4):889–910, 2005.
14. X. L. Wang and Y. Feng. Inter-comparison of extra-tropical cyclone activity in eight reanalysis datasets. *EGU General Assembly Research Abstract*, 2014.
15. M. F. Wehner, K. A. Reed, F. Li, Prabhat, J. Bacmeister, C.-T. Chen, C. Paciorek, P. J. Gleckler, K. R. Sperber, W. D. Collins, A. Gettelman, and C. Jablonowski. The effect of horizontal resolution on simulation quality in the community atmospheric model, cam5. 1. *Journal of Advances in Modeling Earth Systems*, 2014.