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
prabhat@lbl.gov
[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.

The rights of this work are transferred to the extent transferable according to title 17 U.S.C. 105.

© Springer International Publishing Switzerland 2015
G. Azzopardi and N. Petkov (Eds.): CAIP 2015, Part II, LNCS 9257, pp. 426–436, 2015.
DOI: 10.1007/978-3-319-23117-4_37

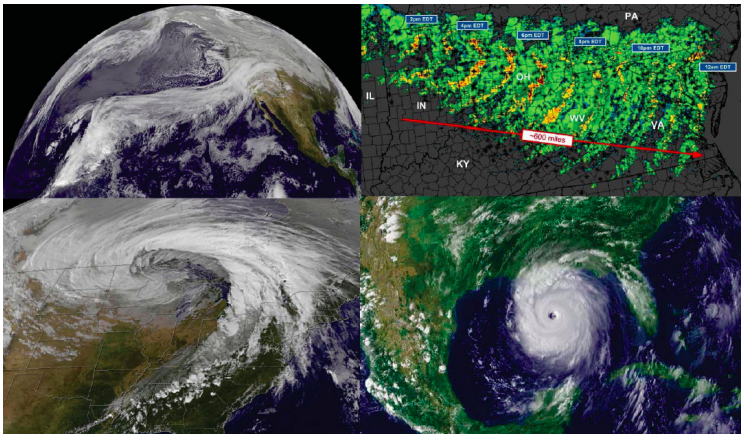

Fig. 1. Examples of extreme weather phenomena observed through satellite and radar. Clockwise from bottom-left: Extra-Tropical Cyclone, Atmospheric River, Derecho and Tropical Cyclone events.

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 analysis 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