Big Data in Climate: Opportunities and Challenges for Machine Learning

Anuj Karpatne
Department of Computer Science and Engineering,
University of Minnesota
karpa009@umn.edu

Vipin Kumar
Department of Computer Science and Engineering,
University of Minnesota
kumar001@umn.edu

ABSTRACT

The climate and Earth sciences have recently undergone a rapid transformation from a data-poor to a data-rich environment. In particular, massive amount of data about Earth and its environment is now continuously being generated by a large number of Earth observing satellites as well as physics-based earth system models running on large-scale computational platforms. These massive and information-rich datasets offer huge potential for understanding how the Earth's climate and ecosystem have been changing and how they are being impacted by humans actions. We discuss the challenges involved in analyzing these massive data sets as well as opportunities they present for both advancing machine learning as well as the science of climate change.

KEYWORDS

Climate Science, Earth Observation Data, Machine Learning

1 OPPORTUNITIES FOR BIG DATA

Climate science has experienced a rapid transformation from a data-poor to a data-rich phase in the last few decades, with data from Earth-observing satellites launched by organizations such as NASA, SpaceX, and European Space Agency (ESA), and massive volumes of data from model simulations that are being generated by multiple groups of climate scientists across the world. The growing size and richness of climate data provide numerous opportunities for data science to improve our understanding of the Earths climate. They also provide answers to some of the pressing questions related to climate change mitigation and adaptation [18, 19].

First, data science methods can play a major role in discovering key climatic processes such as teleconnections, which represent pairs of distant regions in the world that show coupled climate activity. A well-known example of such phenomena is the El-Nino Southern Oscillation in the West Pacific Ocean. Automated discovery of teleconnections using data science methods (e.g., recent network-based algorithms [12]) can help us discover previously unknown phenomena in climate[15, 16]. We can also use data science methods to identify relationships in climate science that exist beyond pairs of regions such as tripoles [3, 14]. Insights gained from such analyses can also help in evaluating and refining climate models based on their ability to reproduce vital climatic processes.

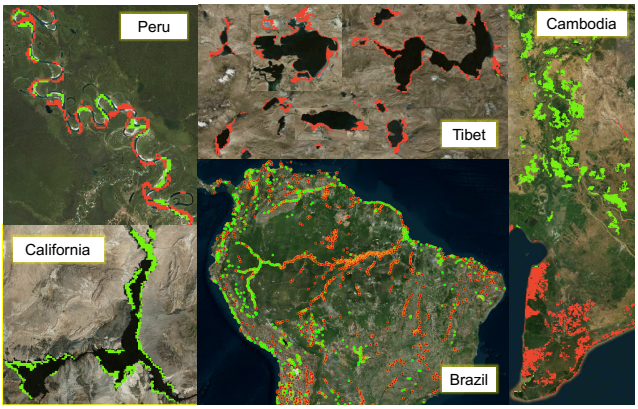

Figure 1: Examples of changes identified by our global surface water monitoring system (green indicates loss of water and red indicates gain in water in the last 15 years). This includes river migrations in floodplains of Peru (possibly aided by increased soil erosion due to deforestation in the Amazonian tropical forests), expanding glacial lakes in Tibet due to melting glaciers, declining water supplies in drought-stricken lakes of California and Cambodia, and increasing constructions of dams and reservoirs in Brazil and around the world that have a variety of ecological impacts. (Image backgrounds are courtesy of Bing Aerial Imagery.)

Second, the vast amount of remote sensing data being collected by Earth-observing satellites can help us monitor critical environmental resources and their interactions with the changing climate. Some examples include monitoring the dynamics of surface water bodies [2, 13] that are impacted by changing climate and human actions (Figure **??**), mapping tropical forest fires [1, 17] that are one of the major contributors of greenhouse gas emissions worldwide, monitoring conversions of tropical forests to oil palm plantations [8–10] and understanding how extreme rainfall patterns are impacted by climate change [6].

KDD '17, August 13-17, 2017, Halifax, NS, Canada

ACM ISBN 978-1-4503-4887-4/17/08.
https://doi.org/10.1145/3097983.3105810

2 CHALLENGES AND RESEARCH NEEDS

Although big data in climate offers numerous research opportunities and the data science community is increasingly becoming eager to explore applications in climate domains [7], there are a number of challenges in utilizing the full potential in climate data for accelerating scientific discovery, relative to the level of success achieved by data science in the commercial arena.

One challenge is that, while traditional data science algorithms are designed for handling well-defined objects, such as items bought in market-basket transactions or lists of friends in social networks, objects of interest in climate science often appear as loosely defined patterns in continuous space-time fields that evolve over space and time. For example, ocean eddies (swirling bodies of water and nutrients in the ocean) appear as changes in sea surface height data with loose boundaries around their edges.

Another challenge is that climate science problems often involve the complex nature of relationships among physical variables that are difficult to extract from the limited number of climate records. For example, high-quality sensor measurements of climate variables on a global scale are only available for the recent past (40 to 100 years). This limits the usefulness of several state-of-the-art data science algorithms such as deep learning, whose success in speech and image recognition problems have been greatly enabled by the internet-scale availability of data in these domains. In fact, black-box data science methods, that are oblivious to the rich understanding of the physical processes driving climatic phenomena, have met with limited success in climate science [4].

To fully capitalize the power of big data for accelerating scientific discovery in the domain of climate, there is an increasing interest in developing a systematic way of integrating climate science knowledge in state-of-the-art data science algorithms. This theory-guided data science paradigm [5, 11] is expected to be a key enabler in advancing our knowledge of the Earths climate system and informing adaptation and mitigation policies related to combating climate change.

3 ACKNOWLEDGMENTS

This work was supported by NSF grant IIS-1029771 and NASA awards 14-CMAC14-0010 and NNX12AP37G.

REFERENCES

[1] 2017. Global Burned Area Mapping using Satellite Data. https://z.umn.edu/fireviewer. (2017). [Online; accessed 14-July-2017].
[2] 2017. Global Surface Water Monitoring System. http://z.umn.edu/monitoringwater. (2017). [Online; accessed 14-July-2017].
[3] Saurabh Agrawal, Gowtham Atluri, Anuj Karpatne, William Haltom, Stefan Liess, Snigdhansu Chatterjee, and Vipin Kumar. 2017. Tripoles: A New Class of Relationships in Time Series Data. In *Knowledge Discovery and Data Mining*. ACM.
[4] Peter M Caldwell, Christopher S Bretherton, Mark D Zelinka, Stephen A Klein, Benjamin D Santer, and Benjamin M Sanderson. 2014. Statistical significance of climate sensitivity predictors obtained by data mining. *Geophysical Research Letters* 41, 5 (2014), 1803–1808.
[5] James H Faghmous and Vipin Kumar. 2014. A big data guide to understanding climate change: The case for theory-guided data science. *Big data* 2, 3 (2014), 155–163.
[6] Subimal Ghosh, Debasish Das, Shih-Chieh Kao, and Auroop R Ganguly. 2012. Lack of uniform trends but increasing spatial variability in observed Indian rainfall extremes. *Nature Climate Change* 2, 2 (2012), 86–91.
[7] Climate Informatics. 2017. International Conference on Climate Informatics. http://climateinformatics.org. (2017). [Online; accessed 14-June-2017].
[8] Xiaowei Jia, Ankush Khandelwal, James Gerber, Kimberly Carlson, Paul West, and Vipin Kumar. 2016. Learning large-scale plantation mapping from imperfect annotators. In *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 1192–1201.
[9] Xiaowei Jia, Ankush Khandelwal, N Guru, James Gerber, Kimberly Carlson, Paul West, and Vipin Kumar. 2017. Predict land covers with transition modeling and incremental learning. In *SIAM International Conference on Data Mining*. SIAM.
[10] Xiaowei Jia, Ankush Khandelwal, Guruprasad Nayak, James Gerber, Kimberly Carlson, Paul West, and Vipin Kumar. 2017. Incremental Dual-memory LSTM in Land Cover Prediction. In *Knowledge Discovery and Data Mining*. ACM.
[11] Anuj Karpatne, Gowtham Atluri, James Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. 2017. Theory-guided data science: A new paradigm for scientific discovery. In *IEEE Transactions on Knowledge and Data Engineering, arXiv: 1612.08544*.
[12] Jaya Kawale, Stefan Liess, Arjun Kumar, Michael Steinbach, Peter Snyder, Vipin Kumar, Auroop R Ganguly, Nagiza F Samatova, and Fredrick Semazzi. 2013. A graph-based approach to find teleconnections in climate data. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 6, 3 (2013), 158–179.
[13] Ankush Khandelwal, Anuj Karpatne, Miriam Marlier, Julia Kim, Dennis Lettenmaier, and Vipin Kumar. 2017. An Approach for Global Monitoring of Surface Water Extent Variations Using MODIS Data. In *Remote Sensing of Environment*.
[14] Stefan Liess, Saurabh Agrawal, Snigdhansu Chatterjee, and Vipin Kumar. 2017. A Teleconnection between the West Siberian Plain and the ENSO Region. *Journal of Climate* 30, 1 (2017), 301–315.
[15] Stefan Liess, Arjun Kumar, Peter K Snyder, Jaya Kawale, Karsten Steinhaeuser, Frederick HM Semazzi, Auroop R Ganguly, Nagiza F Samatova, and Vipin Kumar. 2014. Different modes of variability over the Tasman Sea: Implications for regional climate. *Journal of Climate* 27, 22 (2014), 8466–8486.
[16] Mengqian Lu, Upmanu Lall, Jaya Kawale, Stefan Liess, and Vipin Kumar. 2016. Exploring the Predictability of 30-Day Extreme Precipitation Occurrence Using a Global SST–SLP Correlation Network. *Journal of Climate* 29, 3 (2016), 1013–1029.
[17] Varun Mithal, Guruprasad Nayak, Vipin Kumar, Ankush Khandelwal, Nikunj C. Oza, and Rama Nemani. 2017. RAPT: Rare Class Prediction in Absence of True Labels. In *Transactions on Knowledge and Data Engineering*.
[18] University of Minnesota. 2017. Understanding Climate Change: A Data Driven Approach. http://climatechange.cs.umn.edu/. (2017). [Online; accessed 14-June-2017].
[19] National Science Foundation Report. 2016. Using data to better understand climate change. https://nsf.gov/discoveries/disc_summ.jsp?cntn_id=189519. (2016). [Online; accessed 14-July-2017].