7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

# A VISION FOR THE DEVELOPMENT OF BENCHMARKS TO BRIDGE GEOSCIENCE AND DATA SCIENCE

Imme Ebert-Uphoff[1], David R. Thompson[2], Ibrahim Demir[3], Yulia R. Gel[4], Mary C. Hill[5], Anuj Karpatne[6], Mariana Guereque[7], Vipin Kumar[6], Enrique Cabral-Cano[8], Padhraic Smyth[9].

***Abstract*—**

**The massive surge in the amount of observational field data demands richer and more meaningful collaboration between data scientists and geoscientists. This document was written by members of the Working Group on Case Studies of the NSF-funded RCN on *Intelligent Systems Research To Support Geosciences (IS-GEO, https://is-geo.org/)* to describe our vision to build and enhance such collaboration through the use of specially-designed benchmark datasets. Benchmark datasets serve as summary descriptions of problem areas, providing a simple interface between disciplines without requiring extensive background knowledge. Benchmark data intend to address a number of overarching goals. First, they are concrete, identifiable, and public, which results in a natural coordination of research efforts across multiple disciplines and institutions. Second, they provide multi-fold opportunities for objective comparison of various algorithms in terms of computational costs, accuracy, utility and other measurable standards, to address a particular question in geoscience. Third, as materials for education, the benchmark data cultivate future human capital and interest in geoscience problems and data science methods. Finally, a concerted effort to produce and publish benchmarks has the potential to spur the development of new data science methods, while providing deeper insights into many fundamental problems in modern geosciences. That is, similarly to the critical role the genomic and molecular biology data archives serve in facilitating the field of bioinformatics, we expect that the proposed geosciences data repository will serve as "catalysts" for the new discicpline of geoinformatics. We describe specifications of a high quality geoscience benchmark dataset and discuss some of our first benchmark efforts. We invite the Climate Informatics community to join us in creating additional benchmarks that aim to address important climate science problems.**

Corresponding author: Imme Ebert-Uphoff, iebert@colostate.edu. [1]Electrical & Computer Engineering, Colorado State University, Fort Collins, CO. [2]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA. [3]Civil & Environmental Engineering, University of Iowa, Iowa City, IA. [4]Mathematical Sciences, University of Texas at Dallas, Richardson, TX. [5]Geology, University of Kansas, Lawrence, KS. [6]Computer Science & Engineering, University of Minnesota, Minneapolis, MN. [7]Geological Sciences, University of Texas, El Paso, TX. [8]Instituto de Geofisica, Universidad Nacional Autónoma de México, Coyoacán, CDMX, Mexico. [9]Computer Science, University of California, Irvine, CA..

## I. MOTIVATION

For decades there has been a strong trend in the geosciences in the direction of larger, more diverse datasets that demand sophisticated mathematical and computer science expertise [1]. This is a consequence of improvements in computing power, which permit far more sophisticated physical modeling; improvements in measurement technology, which permit acquisition of high-resolution large-scale datasets; and demands of challenging problems such as measuring the planet's response to a changing climate. Meeting these challenges requires rich communication between the geoscientists familiar with the application domain and data scientists that could bring novel computational methods to the field. Closing this gap is a primary goal of the Climate Informatics workshop series, and is shared by this benchmark development effort. This document describes the manner in which benchmark standard datasets (or simply, benchmarks) of typical geoscience data analysis problems can bridge the two communities.

### A. Relation to Existing Efforts

Benchmarks can be seen as an extension of classic data repositories. In particular, classic data repositories, such as those maintained by *NCAR* [2], *NOAA* [3], *NASA* [4], *USGS* [5], and related repositories [6], [7], provide vast amounts of data, but only domain scientists would know how to use the data efficiently, which questions to ask, and how to set up an interesting analysis [8]. The same holds for repositories maintained by journals, such as the *Geoscience Data Journal* [9] and *Nature Scientific Data* [10]. We seek to bridge the gap between the geoscientist and the data scientist by having geoscientists preselect and preprocess interesting data, couple them with interesting and unsolved science questions, and add data documentation and background explanations suitable for non-domain scientists. The key to the benchmarks is this *packaging* of existing data with science questions suitable for data scientists.

The proposed benchmarks can also be seen as an extension of existing efforts originating from the data science community, such as the CI Hackathon events [11], [12], the UIOWA Midwest Big Data Hackathon

[13], the Challenges in Machine learning events [14], the Kaggle data science platform [15], and the UCI Machine Learning Repository [16]. (In fact, we may incorporate the CI 2016 Hackathon topic [12] and topics from the UIOWA Midwest Big Data Hackathon [13] as benchmarks.) The following aspects distinguish our benchmark datasets from such data science competitions: (1) benchmarks tend to be more open-ended, i.e. there might be no pre-defined performance measure; (2) benchmark data sets are meant to initiate and stimulate interdisciplinary discussion, and in turn to faciliate *long-term* collaborations between data scientists and domain scientists; (3) the benchmark data in their current form do not aim to focus on comparison among different participant groups (i.e., no competition); however, the benchmarks can also be utilized for various data science contests and challenges involving analysis, modeling, validation, and prediction.

### B. Specific Goals

The benchmarks are intended to serve several goals, including:

(1) A **means for two-way communication** to connect the two disciplines, geoscience and data science: (i) data scientists learn about typical data analysis tasks in the geosciences, including typical properties of geoscience data, the types of science questions geoscientists are interested in, and existing approaches for data analysis in the geosciences; (ii) geoscientists learn about potential new methods, tools, and services for data analysis.

(2) Benchmarks seek to **stimulate new collaborations**, which may lead to discovery of new approaches and methods for data analysis in the geosciences; science advances by gaining new insights from geo data using the new approaches; formation of new collaboration teams that will work together in the future.

(3) A **permanent repository of complex data sets, representative of geoscience problems** is an important resource for education, research, and to promote an emergent coordination of research activities and conversations in the research literature that build off each other (not possible when every lab has an independent data set with its own idiosyncrasies).

## II. Desired Benchmark Characteristics

Given the list of goals in the preceding section, what are the key characteristics and elements of an ideal benchmark set? Below is a list of properties that we believe make a data set particularly suitable as a benchmark in this context. An outstanding benchmark is expected to satisfy many, but usually not all of these characteristics.

**High Impact:** A problem with high potential impact should be chosen, and the connection to that impact clearly spelled out, namely how will the proposed tasks contribute to advances in science or benefit society?

**Active Geoscience research area:** To stimulate long-term interactions between geoscientists and data scientists, the benchmark should come from an active research area, i.e. a group of geoscientists should be eager to continue to work on the topic, to answer questions from the data scientist(s), and to help him/her interpret any analysis results.

**Challenge for data science:** The problem should be challenging for the data scientists in some way. This is almost a given for geoscience applications, because 1) if the analysis was straightforward the geoscientist would have done the analysis him/herself; 2) geoscience data by their very nature tend to pose several challenges for standard data analysis methods (see Section II-A below). Known challenges should be spelled out for each benchmark, but some challenges will only become apparent during the analysis.

**Data science generality and versatility:** Ideally, solutions generated from the data set analysis and proposed models and algorithms will not only help to address a stated set of problems in geosciences, but also will be applicable to a broad range of other settings and possibly other disciplines, i.e. will stimulate and facilitate development of new methodology in data science.

**Rich information content:** Ideally the data set provides stimulus for analysis at many different levels, i.e. it lends itself to answering more than one science question. If so, one can gain a lot from a single data set.

**Hierarchical problem statement:** Each benchmark should include a data set and a clear description of what types of analyses are suggested. Ideally, there is a hierarchy of analysis tasks, ranging from relatively straight-forward tasks to more open-ended tasks.

**A means for evaluating success:** Data scientists need some kind of means to evaluate whether their algorithms are successful in solving the problem. Ideally, some kind of performance measure should thus be included for at least some of the tasks. However, in very open-ended applications, the performance measure might be developed during the collaboration.

**Quick start guide:** It should be as easy as possible for data scientists to start working with the data. Data scientists focus on data first, so the data needs to be easily accessible, and ideally there would be quick-start instructions on how to explore them. We seek to include for each benchmark a data use tutorial, consisting of (1) code snippets in a well-known framework (e.g., Matlab, Python, R) that illustrate how to read and visualize the data, potentially also illustrating some sample analysis steps; (2) plots generated from the code snippets that illustrate some of the data properties, so that data scientists can get a better feeling for the problem before even touching the data.

**Understandable geoscience context:** Geoscience data generally has a rich background, ranging from the motivation for collecting the data in the first place (science question), to the instruments used to take it,

the pre-processing that has already taken place, and the science questions it seeks to answer. Providing a brief summary of this background, in a way that is easy for someone outside the field to understand (no jargon), results in more efficient collaboration and may yield a more meaningful analysis of the data.

**Citability:** As discussed in the article on the *Geoscience Paper of the Future* [17], it is crucial to provide for each data set (1) a license specifying conditions for use, and (2) a unique and persistent identifier to make it citable, and later allow search engines to easily find all research papers using it. Both criteria can be met by using *Zenodo* (https://zenodo.org/) to host the data sets. Zenodo is a data repository run by CERN, that provides free hosting of data sets up to 50 GB, provides a selection of license terms and assigns a unique DOI number to each data set.

**Communication between researchers:** A public Google document provides both an FAQ and a communication channel for the domain experts and anyone working with the data. Researchers may use it to ask questions, exchange experiences and discuss results.

### A. Suitable data science methods

Many data science methods cannot be directly applied to geoscience data, because of the challenging properties of such data. Karpatne et al. [18] categorize the most important challenging properties as follows: spatiotemporal structure; high dimensionality; heterogeneity in space and time; existence of objects with amorphous spatial/temporal boundaries; multi-scale/multi-resolution data; low sample size; paucity or absence of ground truth; noise, incompleteness, and uncertainty in data. It will be useful to identify for each benchmark which challenging properties are present.

Furthermore, it is difficult to convince geoscientists to use any method they do not understand. In fact geoscientists strongly prefer *transparent* methods, which allow them to follow the basic reasoning and generate novel scientific insights, over *black box* methods [19].

### B. Distribution and Advertisement of benchmarks

All benchmarks will be featured on the IS-GEO website, and advertised through papers (such as this one), talks, and mailing lists, and we will reach out personally to data scientists through the IS-GEO members to make them aware of these benchmarks.

## III. Sample Benchmarks

The IS-GEO benchmark project was created in Spring 2017. To date we have created one benchmark and are working on two more.

The first benchmark was developed in collaboration with researchers at the Jet Propulsion Laboratory (JPL), and deals with the automatic analysis of their imaging spectrometer data in order to detect significant sources of methane in the atmosphere [20], [21]. Methane (CH4) is a powerful Greenhouse Gas in the atmosphere and it is essential to determine its most important sources in the environment, such as geologic seeps, animal husbandry, decomposition in landfills, and oil and gas extraction and production. The ultimate goal of this benchmark is to develop methods for the reliable detection, and potentially classification, of methane sources from imaging spectrometer data. The key challenge is to distinguish methane sources from background noise in the spectrometer images. Domain experts currently perform this task manually by visual inspection of the imaging spectrometer data.

With regard to the requirements from Section II, this benchmark satisfies many of them. Namely, it is a high impact application, as it has the potential to reduce greenhouse gas emissions, and thus global warming; it is an active research area of research institutions such as JPL; it provides a rich, multi-layered and challenging playground for data scientists, because the data includes high levels of noise, as well as artifacts from roads and buildings that may be addressed through a variety of sophisticated statistical and image processing techniques; the problem statement consist of a hierarchy of tasks of increasing difficulty; we developed a quick start tutorial with Matlab code examples and visualization of sample data; there are manually labeled results for methane detection that can be used to evaluate performance for the simpler tasks, while the remaining tasks are more open-ended.

We are currently working with the team of the 2016 Climate Informatics Hackathon event to extend their challenge, prediction of sea ice cover based on several atmospheric variables [12], to a benchmark. We also collected a list from the IS-GEO community containing ten additional benchmark topics to consider.

## IV. An Invitation to the CI Community

We invite the members of the Climate Informatics community to get involved in this effort. In particular, we would appreciate feedback on the general vision presented here, and any collaboration for the creation of additional benchmarks. Furthermore, we invite you to find out more about the general IS-GEO initiative (https://is-geo.org/) and to become a member of that community as well.

## V. Acknowledgments

We are grateful to the ChaLearn organization for sharing resources and giving helpful advice. Their guidelines for setting up Challenges in Machine Learning [14] served as a great starting point. This activity is part of the IS-GEO Research Collaboration Network funded by the NSF (Award #1632211, EarthCube RCN IS-GEO: Intelligent Systems Research to Support Geosciences).

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

## REFERENCES

[1] I. Demir, H. Conover, W. F. Krajewski, B.-C. Seo, R. Goska, Y. He, M. F. McEniry, S. J. Graves, and W. Petersen, "Data-enabled field experiment planning, management, and research using cyberinfrastructure," *Journal of Hydrometeorology*, vol. 16, no. 3, pp. 1155–1170, 2015.

[2] National Center for Atmospheric Research (NCAR), "NCAR Community Data Portal (CDP)." http://cdp.ucar.edu/.

[3] National Oceanic and Atmospheric Administration (NOAA), "National Centers for Environmental Information (NCEI) - Data Access." https://www.ncdc.noaa.gov/data-access.

[4] National Aeronautics and Space Administration (NASA), "NASA's Open Data Portal." https://data.nasa.gov/.

[5] U.S. Geological Survey , "U.S. Geological Survey Science Data Catalog." https://data.usgs.gov.

[6] Interdisciplinary Earth Data Alliance (IEDA), "Observational solid earth data from the ocean, earth, and polar sciences - data repositories." http://app.iedadata.org/compliance/dmp/replist.php.

[7] U.S. Government, "The home of the U.S. Governments open data." https://www.data.gov/.

[8] I. Ebert-Uphoff and Y. Deng, "Three steps to successful collaboration with data scientists," *Earth and Space Science (EOS)*, vol. 98, 2017. https://doi.org/10.1029/2017EO079977.

[9] Royal Meteorological Society, "Geoscience data journal." An Open Access Journal, published by Wiley. http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)2049-6060.

[10] Macmillan Publishers (part of Springer Nature), "Scientific Data." An open-access journal for descriptions of scientifically valuable datasets, and research that advances the sharing and reuse of scientific data. https://www.nature.com/sdata/.

[11] Paris Saclay Center for Data Science, "Climate informatics hackathon 2015," 2015. https://www.lri.fr/~kegl/Ramps/edaElNino.html.

[12] B. Kégl and A. Rhines, "Climate informatics hackathon 2016," in *Proceedings of the 6th International Workshop on Climate Informatics (CI 2016)*, 2016. http://dx.doi.org/10.5065/D6K072N6.

[13] University of Iowa, "UIOWA Midwest Big Data Hackathon." http://bigdata.uiowa.edu.

[14] ChaLearn, "Challenges in machine learning." http://www.chalearn.org/.

[15] Kaggle, "Kaggle: Your home for data science." https://www.kaggle.com/datasets.

[16] UC Irvine, Center for Machine Learning and Intelligent Systems, "UCI machine learning repository." https://archive.ics.uci.edu/ml/datasets.html.

[17] Y. Gil, C. H. David, I. Demir, B. T. Essawy, R. W. Fulweiler, J. L. Goodall, L. Karlstrom, H. Lee, H. J. Mills, J.-H. Oh, *et al.*, "Toward the geoscience paper of the future: Best practices for documenting and sharing research from data to software to provenance," *Earth and Space Science*, vol. 3, no. 10, pp. 388–415, 2016.

[18] A. Karpatne, H. A. Babaie, S. Ravela, V. Kumar, and I. Ebert-Uphoff, "Machine learning for the geosciences - opportunities, challenges, and implications for the ML process," in *Workshop on Mining Big Data in Climate and Environment (MBDCE 2017), 17th SIAM International Conference on Data Mining (SDM 2017)*, April 2017.

[19] A. Karpatne, G. Atluri, J. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar, "Theory-guided data science: A new paradigm for scientific discovery from data," *IEEE Transactions on Knowledge and Data Engineering*, 2017.

[20] C. Frankenberg, A. K. Thorpe, D. R. Thompson, G. Hulley, E. A. Kort, N. Vance, J. Borchardt, T. Krings, K. Gerilowski, C. Sweeney, *et al.*, "Airborne methane remote measurements reveal heavy-tail flux distribution in four corners region," *Proceedings of the National Academy of Sciences*, p. 201605617, 2016.

[21] D. Thompson, I. Leifer, H. Bovensmann, M. Eastwood, M. Fladeland, C. Frankenberg, K. Gerilowski, R. Green, S. Kratwurst, T. Krings, *et al.*, "Real-time remote detection and measurement for airborne imaging spectroscopy: a case study with methane," *Atmospheric Measurement Techniques*, vol. 8, no. 10, pp. 4383–4397, 2015.