US CLIVAR Working Group "Emerging Data Science Tools for Climate Variability and Predictability"

1. Motivation

The CLIVAR research community studies climate variability and change on a range of time scales, through modeling and the analysis of observations. Both sources of information have exploded in size and complexity over the past decade, and as is the case in other sciences, new computational tools have been and are being developed to meet this challenge. New data analysis methodologies and algorithms that undergird these tools are the province of what we refer to here as the "data sciences": computer science and statistics, with machine learning straddling the two. The data sciences have seen rapid development in the past decade and is transforming many areas of engineering, science, and the global economy (e.g., McAfee and Brynjolfsson 2017; Lubberts and Miikkulainen 2001; Silver et al. 2016; Khan et al. 2001; Zhou et al. 2002; Karabatak and Ince 2009; Hinton et al. 2012a,b; Dahl et al. 2011; LeCun et al. 2015; Silver et al. 2016; Tao et al. 2016). New tools from the data sciences may allow, for example, the discovery of relationships and processes in large data sets that may hitherto have gone unnoticed, or provide computationally efficient emulation of physical models. At the same time, new tools from the data sciences can foster a virtuous cycle in which new data directly inform models, which in turn inform what data to acquire.

We propose a US CLIVAR working group whose main aim is to help foster the understanding, adoption and further development of modern data science tools for the analysis of large-to-massive climate data sets. It brings together experts in Earth science, statisticians, and computer scientists with the specific goal of fostering collaboration across these disciplinary boundaries to achieve CLIVAR's scientific objectives. This overarching goal intentionally breaks tradition with past CLIVAR working groups, which have focused on processes or predictability opportunities emerging primarily *within* the Earth Sciences disciplines. Yet, at this time of rapid growth in the data sciences, we feel it is appropriate for a different sort of WG to form – with a mandate to identify consensus where it exists within a nascent part of our community that is experimenting with powerful new methods, some of which originated *outside* the Earth Sciences.

While there are many directions this WG could take, touching on nearly all aspects of science covered by CLIVAR, we choose to focus this WG on three specific themes:

1. How should we change modeling practice in areas that are becoming data-rich (e.g. global cloud-resolving simulation) or where physical constraints on models are weak (e.g., land and ocean biosphere models)? In other words -- what is the outlook for actually replacing process parameterization with machine learning emulation in these specific arena? This is inspired by a glimmer of recent success in using machine learning to emulate sub-grid cloud-resolving physics (Bretherton and Brenowitz, 2018; O'Gorman et al. 2018; Gentine et al. 2018; Rasp et al. 2019) to build a next generation of data-driven climate models (Schneider et al. 2017). Furthermore, in what way should uncertainties in these algorithms/parameters be incorporated into this new type of parameterization, or used to further improve climate model simulations? Outstanding challenges and opportunities include, for instance, interpretability, stability, stochasticity, out-of-sample generalizability, uncertainty quantification, and philosophical trade-offs with physically-based approaches intrinsic to such an approach.

2. What is the potential for data-driven discovery to identify and substantiate new patterns in climate variability, and to link them to predictability on subseasonal-to-climate timescales? For decades the field has utilized data-driven approaches to identify covariations across the climate system and capture relevant patterns (e.g. empirical orthogonal functions (EOF) analysis). New methods from the data sciences have the potential to offer additional "tools for the toolbox". For example, causal inference theory may be useful for studying teleconnections, or improving empirical and dynamical forecast systems understanding (Runge et al. 2018, Green et al., 2017). Machine learning methods that were historically viewed as black boxes now have the potential to be partially interpreted (i.e. for scientists to learn how the decisions were made) thanks to recent

advances in visualization and interpretibility techniques (i.e. optimal input analysis, layer-wise relevance propagation). Outsanding challenges and opportunities include further developing interpretability methods for climate applications and finding ways to leverage these often data-hungry methods for studying processes and uncertainties with infrequent samples, such as extreme events.

3. Which specific emerging data science tools and methods does current experience suggest are most likely to provide breakthroughs and how should the relevant subcommunity (and this WG) orient itself to maximize the benefit? These questions reflect the fact that we are in the infancy of adopting emerging data science tools in the CLIVAR community, and thus require some scope and flexibility in sharing the collective experience and adapting accordingly. This includes learning from other relevant initiatives that have also begun to use those new methods such as Google, IBM, Vulcan, Climate Corps, the Climate Modeling Alliance, and Jupiter Intelligence.

The proposed WG will bring together domain experts in Earth science, statisticians, data scientists, and applied mathematicians. There have been a number of workshops and programs relevant to the focus of the proposed WG. However, these examples are driven from within their specific communities rather seeking to cover the breadth of the proposed WG. For example:

1. The *Statistical and Applied Mathematical Sciences Institute* (SAMSI) ran a "Program on Mathematical and Statistical Methods for Climate and the Earth System" in 2017-2018 [link].
2. The *Statistical Methods for Atmospheric and Oceanic Sciences (STATMOS) Research Network* [link] is a network primarily of statisticians interested in the climate sciences.
3. The *Society of Industrial and Applied Mathematics* (SIAM) holds a "Mathematics of Planet Earth" conference series, which is broadly on mathematical aspects of the Earth sciences (e.g., dynamical system approaches).
4. *Climate Informatics* [link] holds an annual workshop to bring together researchers from climate science and the areas of statistics, machine learning and data mining to stimulate discussion and foster new collaborations in order to grow the climate informatics community.
5. With support by the Heising-Simons Foundation, Caltech has held a series of workshops in 2017/18 on "Nucleating a Next Generation of Earth System Models," in which data-driven approaches were a focus (see http://workshop.caltech.edu/fesm/).
6. International Meetings on Statistical Climatology (IMSC) organized every three years since the later 1970s to promote good statistical practices in atmospheric and climate sciences (see http://imsc.pacificclimate.org/).

While these activities and others cover particular aspects of the focus of the proposed WG, they do not focus exclusively on the subset of tools and questions most relevant to the CLIVAR community's shared scientific goals. We thus recognize a timely opportunity to build on and link to such efforts, and to focus new discussion on the breadth of topics and subdisciplines brought together in the proposed WG.

2. Objectives, Tasks, Timeline

1) Identifying best practices for interpretability and predictive capacity. While the proposed WG is focused on new tools from the data sciences, many of these tools are historically used as "black boxes". While this can be justified if predictability gains alone are the goal, in climate prediction and variability applications, we also aim to better understand the physical world, and thus, a black box approach alone is not optimal. For example, the predictive success of deep learning is entangled with problems interpreting the results, and with generalization errors and uncertainties when out-of-sample predictions are required, as is often the case in climate predictions. Hence, a key objective of the WG will not only be to explore the use of such tools, but to also investigate to what extent they need to be modified and judiciously combined with existing tools and practices in the climate sciences (e.g., physical modeling, data assimilation), while satisfying physical constraints (energy and mass conservations) to make them useful, all the while quantifying the uncertainties. Meanwhile, we also see the benefit of predictability gains in their own right and will act

to distill best practices based on what subsets of new data science methods are actually proving to have standout skill in statistical climate prediction, even independent of understanding.

2) **Gathering, preparing, and encouraging use of benchmark training data sets** for coordinated methodological intercomparison. Our WG recognizes that those experimenting with data sciences methods in our community are scattered across disciplines and lack curated common reference datasets community based on which to evaluate algorithms and methodological choices. Recognizing that such benchmarks have enabled core advances in data science, for instance in the image recognition community (ImageNet), we see an opportunity to collectively craft attractive requirements for such a dataset that would meaningfully test the prediction skill and interpretability trade-offs of interest to the CLIVAR community. Perhaps a few such datasets already exist in our midst, such as via the Large Ensemble WG, which we plan to coordinate with as one possible reference data source. By also considering this question as an objective, we may begin articulating a well reasoned case for what else should be considered an optimal testbed for our community to exchange methodological notes on.

3) **Providing timely perspectives on emerging data sciences tools** and their potential to advance the field. Artificial intelligence, machine learning, and statistical advances are taking place at such a dizzying pace that it is no coincidence those involved tend to monitor preprints on arXiv as much as the peer-reviewed literature, if only to avoid its lag. In a similar spirit, we envision this WG to act as a rapid clearing house for disseminating those findings and experiences of its members' collective awareness that intersect interestingly with CLIVAR's goals, as they arise. Keeping up is half the battle these days, such that we view this as an objective-worthy community service in its own right. As part of this task we will act to disseminate toolboxes and case examples to help early career scientists.

Tasks

a) Regular webinars: Once every 2 months we will invite 2 speakers (20 minutes each) to present in meetings open to the community. A subsequent 20-minute discussion will be moderated by the WG co-chairs, who will be responsible for collecting data towards Objectives 1) and 2), to be compiled in rough form.

b) Online presence: A github or other natural internet environment will be maintained as a repository of moderated discussion notes and a vehicle for wider dissemination of WG discussions. This will include a "news" module, towards which the webinars in Task a) will conclude with opportunities for all to share any relevant new results on their radar, to inform Objective 3). Meanwhile a dedicated Twitter account for the WG will boost this news and maintain awareness of the WG to early career scientists. WG members will be encouraged to contribute new studies to this repository on a regular basis, at higher frequency than the webinars in Task a)

c) UC-Irvine workshop: In the Spring of 2020 a first community workshop for ~40 participants will be held at the University of California, Irvine under the CLIVAR WG's umbrella. The workshop will conclude with a 2-hour moderated discussion.

d) NASA GISS-Columbia University: in the fall of 2019 a wokshop at the intersection of data sciences and carbon cycle will be held at NASA GISS in New York, NY with about 50 participants from academia but also from the industry.

e) Synthesis stage 1: Near the end of Year 1, a subset of the WG will meet for 1-2 days to arrive at a preliminary assessment of where issues of predictability and interpretability relevant to Objective 1) currently stand, as revealed by Tasks a)-c), and discuss the potential, limitations, and level of broad interest in candidate benchmark datasets uncovered to date.

f) National conference sessions: Beginning in 2020, special sessions will be proposed at relevant conferences such as that of the American Geophysical Union (AGU), American Meteorological Society (AMS), Society of Industrial and Applied Mathematics (SIAM), Joint Statistical Meetings (JSM), etc, to maintain discussion and momentum.

g) Caltech workshop: During Summer 2021, a second community workshop will be held for ~40 participants at Caltech, to be co-sponsored by the Climate Modeling Alliance (CliMA; an alliance of Caltech, JPL, MIT, and the Naval Postgraduate School that is developing a data-driven climate model). This workshop will

focus on how to improve climate models and predictions with data sciences tools, and how these tools may be employed differently in fields such as turbulence modeling, where the underlying equations are known, and biosphere modeling, where detailed process-level understanding is incomplete. CliMA has already committed to covering most of the participant expenses; CLIVAR funding is sought for ~ 20% of the overall workshop, and would be intended to focus exclusively on early career participant support.

h) Synthesis stage 2: Near the end of Year 2, a subset of the WG will meet again for 1-2 days to lay the groundwork for a white paper / review article summarizing workshop and working group findings, as explained more in the next section.

i) Writing stage: During Year 3 we will focus on drafting and writing our capstone manuscript; CLIVAR funding is sought for 100% support of a core team working group meeting (12 members) during this final year to discuss culminating conclusions and legacy steps.

3. Publications and Outreach

Given the growing importance of data science in climate variability and prediction, a strong emphasis will be placed on early career scientists in order to ensure a strong foundation in these emerging techniques to support the future of the field. Thus, the WG will ensure early-career scientists make up a major fraction of participants in events, including workshop attendees, special conference sessions, and contributions to WG papers.

Near the end of Year 1 we will submit a journal review article to BAMS that distills which data science methods have already proved most interesting and useful to the CLIVAR community, and which open questions are in most need of new work, as revealed by consensus among WG members in our kick-off workshop's concluding discussion and regular webinars. At the end of the project a second journal article will synthesize progress made during the course of the WG's tenure, report existing benchmark datasets and itemize agreed best practices for generating ones with maximum utility, and offer an opinion on which unsolved methodological issues are of most importance to the CLIVAR community to resolve.

4. Reporting Plan

The objectives and tasks of the proposed working group are highly relevant to the science goals of US CLIVAR. Data analysis and data science is at the foundation of understanding climate variability and its prediction. In addition, the proposed WG has specific objectives focused on model development and process understanding, thus, this WG would support the missions of all of the US CLIVAR panels (POS, PPAI, PSMI). For this reason, we propose to report our relevant progress to each of the three CLIVAR panels.

5. Leadership and Suggested Membership

Our approach to inclusion and balance is as follows. We have intentionally suggested a young team, with many team members who are within 10 years of receiving their PhD -- early career scientists comprise 75% of our leadership team and 60% of our overall working group members. Female scientists account for half of our travel-supported working group members. In addition to university researchers, NASA, DOE and NOAA members are represented on the core team. We have striven to balance some members with experience and activity at the nexus of machine learning and CLIVAR science with others who have fundamental statistical expertise -- the idea is to help ground conversations about new tools in the context of a broader need to generalize the interpretation of results of ML and other data science techniques.

Leads

1.	† Mike Pritchard (U. California, Irvine)	Convection, multi-scale modeling, ML.
2.	† Elizabeth Barnes (Colorado State U.)	Dynamics, statistics, causal discovery, ML.

Co-Leads

3.	Amy Braverman (NASA JPL)	Statistics, uncertainty quantification, data.
4.	† Pierre Gentine (Columbia U.)	Turbulence, convection, biosphere, ML, causal discovery.

Additional Core Team Members (travel paid)

** denotes Early Career (PhD in last 5 years), † (PhD in last 10 years), ML=Machine Learning*

5.	Richard Smith (U. North Carolina)	Statistical methods in climate sciences.
6.	Imme Ebert-Uphoff (Colorado State U.)	Causal inference, ML
7.	† Ryan Abernathey, (Columbia U.)	Ocean mesoscale modeling, ML.
8.	† Matthew Norman (ORNL)	Numerical algorithms, GPU computing, ML.
9.	† Laure Zanna (NYU)	Ocean mesoscale modeling, ML
10.	*Noah Brenowitz (U. Washington)	Convection, ML.
11.	Bo Li (U. Illinois)	Bayesian hierarchies, climate data.
12.	Emily Kang (U. Cincinnati)	Statistics, data science, observations.

Additional Working Group Members (travel not paid)

13.	Vladimir Krasnopolsky (NOAA EMC)	Predictability, ML.
14.	† Paul Ullrich (UC Davis)	Extreme events, ML.
15.	*Ying Sun (Cornell)	Ecosystem, climate, remote sensing, ML.
16.	† Paul Loikith (Portland State U.)	Statistics, climate variability, observations.
17.	Bruno Sanso (UC Santa Cruz)	Bayesian methods for spatial statistics.
18.	Alicia Karspeck (Jupiter Intelligence)	S2S predictability, ML.

6. Resource Requirements

We plan for this WG to last three years, meeting once per year for 2-3 days, convening community worskhops in years 1 and 2, a capstone core team working group meeting in year 3, holding bimonthly webinars, and publishing two peer reviewed articles. This requires travel funds, especially for early career attendees, and publication fees. Specifically we seek:

- Pub support for a *JAMES* and *EOS* papers.
- Travel support, especially for early career attendees, to attend the two workshops.
- Workshop logistics and travel costs
- Staff support for workshop and telecon organization

References

- Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, 45, 6289– 6298. https://doi.org/10.1029/2018GL078510
- Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2011). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, *20*(1), 30–42. https://doi.org/10.1109/TASL. 2011.2134090
- Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018). Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, 45, 5742– 5751. https://doi.org/10.1029/2018GL078202
- Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A. R., Jaitly, N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97. https://doi.org/10.1109/MSP.2012.2205597
- Hinton, G. E., Srivastava, N., & Krizhevsky, A. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *ArXiv.Org*.
- Karabatak, M., & Ince, M. C. (2009). An expert system for detection of breast cancer based on association rules and neural network. *Expert Systems with Applications*, *36*(2 PART 2), 3465–3469. https://doi.org/10.1016/j.eswa.2008.02.064
- Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, *7*(6), 673–679. https://doi.org/10.1038/89044
- Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539
- Lubberts, A., & Miikkulainen, R. (2001). Co-evolving a go-playing neural network. *Genetic and Evolutionary Computation Conference (Gecco-2001)*, 1–6.
- McAfee and Brynjolfsson, *Machine, Platform, Crowd*: *Harnessing Our Digital Future*, Norton, 2017
- O'Gorman, P. A., & Dwyer, J. G. (2018). Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *Journal of Advances in Modeling Earth Systems*, 10, 2548– 2563. https://doi.org/10.1029/2018MS001351
- Rasp, S., M. S. Pritchard, and P. Gentine, 2018: Deep learning to represent subgrid processes in climate models. *Proc. Natl. Acad. Sci.*, **115**, 9684 LP-9689, doi:10.1073/pnas.1810286115.
- Schneider, T., Lan, S., Stuart, A., & Teixeira, J. (2017). Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. Geophysical Research Letters, 44, 12,396– 12,417. https://doi.org/10.1002/2017GL076101
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*, 484–489. https://doi.org/10.1038/nature16961
- Tao, Y., Gao, X., Ihler, A., Hsu, K., & Sorooshian, S. (2016). Deep neural networks for precipitation estimation from remotely sensed information. *2016 IEEE Congress on Evolutionary Computation, CEC 2016*, 1349–1355. https://doi.org/10.1109/ CEC.2016.7743945
- Zhou, Z. H., Jiang, Y., Yang, Y. B., & Chen, S. F. (2002). Lung cancer cell identification based on artificial neural network ensembles. *Artificial Intelligence in Medicine*, *24*(1), 25–36. https://doi.org/10.1016/S0933-3657(01)00094-X