Breakout Session Report

Data-driven parameterization development using ARM observations and models

Kara Lamb (Columbia University), Peter Jan van Leeuwen (Colorado State University/University of Reading), Po-Lun Ma (PNNL), Marcus van Lier Walqui (Columbia University/NASA GISS)

Description: Data-driven parameterization development for convection, clouds, and aerosol processes in both process level atmospheric models and larger scale Earth System Models has become increasingly popular in recent years. These methods hold significant potential to reduce both structural and parameter uncertainty in physical models and to improve the consistency of the representation of these processes across spatial and temporal scales. In this session we aim to bring together DOE ARM/ASR researchers who are working on atmospheric model parameterization development from a data-driven perspective. In particular, we are interested in discussing:

- Recent work and advances related to atmospheric model parameterization development from a data driven perspective (machine learning, reduced order modeling, Bayesian methods, causal discovery, and data assimilation)
- Data assimilation for parameter estimation, studies focused on improving process level understanding integrating DOE ARM/ASR observations
- Improving the consistency of parameterizations across spatial and temporal scales - identification of processes that are the most important

We would plan to invite a few speakers to motivate the discussion, but will mainly encourage interested speakers from the community to highlight their recent work and advances in this session. We plan to have two sub-sessions, one focused on larger scale models and another on process level studies. Each sub-session would begin with an invited speaker, which would be followed by contributed talks from community members. Following these two sessions, we would plan to have a 30 minute discussion, around the following topics:

- What are best practices? What are common challenges and barriers to advancement?
- Are there community tools and data sources that we should consider developing or that already exist?
- How can these methods help to identify ARM observation gaps?
- Are there future research directions the community can identify and recommend?

Schedule:

Tuesday, August 8, 2023

	Speaker	**Topic**	
2:00 - 2:10	P.J. van Leeuwen	Intro	
2:10 - 2:20	Po-Lun Ma	EAGLES Parameterization development	Remote
2:20 -	H. Morrison	Microphysics Scheme Development	

2:30			
2:30 - 2:40	I. Silber	Arctic Cloud-Base Ice Precipitation Properties Retrieved Using a Bayesian Inference Method	
2:40 - 2:50	N. Riemer	Recent Developments with PART-MC	
2:50 - 3:00	Discussion		
3:00 - 3:10	K. Lamb	Reducing Structural Uncertainty in Microphysical Models	
3:10 - 3:20	M. van Lier Walqui	Perturbed parameter estimation with LES to constrain warm rain microphysical process rates	
3:20 - 3:30	M. DeCaria	Novel Nonlinear Causal Discovery for Strongly Coupled Cloud Systems	
3:30 - 3:40	P. Garg	Exploring the Causal Relationship between Environment, Cloud, Aerosol, and Precipitation Properties using ARM Observations and Machine Learning	Remote
3:40 - 3:50	A. Geiss	Discovering new representations of near-surface momentum and energy exchange using symbolic regression	Remote
3:50 - 4:00	Discussion		

Number of Attendees: ~60 in person; ~20 remote

Summary Authors: Kara Lamb, Peter Jan van Leeuwen, Marcus van Lier Walqui
Po-Lun Ma

Main Discussion

- We had a longer discussion on how to use *machine learning* in ARS-funded research. It is the experience of several attendants that it is important to bring physical insight and intuition into the architecture of the machine. The trial-and-error approaches tend to be less fruitful due to the complexity of the problems we face:
  - It was reported that typically it can take several years of collaboration between atmospheric scientists and machine learners to be able to generate an optimal architecture for a specific task, in this case generating parameterizations based on detailed aerosol particle simulations in an LES model. There was also some discussion about whether there was value in purely data-driven approaches, and many researchers showed a preference for augmenting existing parameterizations with machine learning approaches where there is already significant domain knowledge available.
  - As another example, the EAGLES project found that it is easier and the machine is more accurate when the parameters of an existing physically-based parameterization scheme are estimated, then estimating the full parameterization scheme.
  - While it is good practice to start from a known ML solution to move from an easier to a more complex problem, this does not work well in all cases. For instance, experience

has shown that when moving from 1-D radiative transfer to 3-D radiative transfer, one should not start at the 1-D solution and add more complexity to the machine. Instead, insight into the 3-D radiative transfer shows that the 1-D solution is less relevant, and it is better to start from the 1-D direct beam and expand to 3D from that solution.

- *Parameter estimation* is a highly nonlinear problem, enforcing us to use nonlinear estimation methods such as Markov-Chain Monte-Carlo sampling. However, this method needs millions of samples, corresponding to millions of model runs. This is not possible for high-dimensional models. However, one can use machine learning to generate much faster surrogate models, and exciting new work in this area was presented.
- Machine learning, and specifically methods like symbolic regression can be used to *learn equations from time-series observations*, an example being AI Feynman. It was shown in a turbulent surface boundary layer example that the resulting equations can fit the ARM observations remarkably well, better than existing turbulence models. However, it is not easy to understand the resulting equation. The path forward might be to stratify the data further, e.g. according to fetch length, and, if successful, that would imply that fetch length is an important input variable in the symbolic regression. *Reduced order modeling* methods can also provide insights into the number of independent degrees of freedom that need to be parameterized to accurately reproduce higher fidelity models, as was demonstrated for microphysics parameterizations.
- The *causal discovery* talks triggered a discussion on how to handle different time scales. The issue is that while some processes act on time scales of minutes or less, others influence a cloud system over time scales of days. How does one combine drivers acting on these very different time scales in one causal web? This is an issue that needs further study. It is well-known that if the time resolution of the observations is larger than the interaction time scale of the variables involved, the causal attribution can be in error, simply because important processes can be missed completely. Hence, simply smoothing variables to longer time scales will not help. Two solutions were discussed, chaining causal webs for short time scales, and using information from fast time-scale causal webs to inform long-time-scale causal webs, but it was clear that both have issues.

Key Findings

Issues

- The discussion brought up a number of issues and questions about best practices for implementing these different approaches for specific science problems, as detailed above. Many of these points would benefit from *significantly longer discussion and sharing of experience*, particularly as these methods become increasingly more prevalent within the ASR/ARM community.
- There are tradeoffs in terms of resource allocation between exploring new ideas from the rapidly developing data science and scientific machine learning community, and focusing on developing parameterizations that can be implemented in models in a relatively short time period. This might be described as an “exploration-exploitation” trade-off, but it can be detrimental to not have a balanced approach to allow for the cross-fertilization of new and very innovative approaches that may allow for significant advances, while also not losing focus on the ultimate goal of improving process-level understanding and modeling efforts for the atmosphere.

Needs

- Given that many researchers in this area are confronting similar challenges when it comes to directly implementing ML parameterizations in models, one need that was brought up was the awareness of currently available tools, e.g. to implement ML codes in models. Is this a possible opportunity for *DOE-funded/advised open-source software development*?
- We note that there was significant community interest in this topic, and also a need for greater knowledge and knowledge-sharing among researchers of these methods. Many participants expressed interest in continuing the discussion beyond the amount of time we were able to allot during the breakout session.

Decisions

N/A

Future Plans

- Many of the critical issues of interest in our session were also discussed heavily in the kilometer-scale modeling section, e.g. how to best do model-observation comparison, evaluation and constraint. In general these issues constitute the obstacles and opportunities involved in bridging between ARM/ASR and DOE modeling efforts (E3SM, ESMD, etc.).
- After the breakout session, we discussed whether addressing these questions in *a targeted workshop*, which would allow greater time for discussion and sharing of knowledge from community members, would be the logical next step. Longer term, we suggest that organizing all these issues around *a new working group* (that bridges traditionally separate scientific research areas) might be appropriate.

Action Items

- Discuss with program managers a *targeted workshop on model-observation comparison, and how to use observations to improve models*. This workshop should include data-driven parameterization development, and issues addressed in the kilometer-scale modeling session (in particular those related to observations and how to use observations to improve these models).