Pacific Northwest
NATIONAL LABORATORY

PNNL-30561

Developing a simulator-based satellite dataset for using machine learning techniques to derive aerosol-cloud-precipitation interactions in models and observations in a consistent framework

September 2020

Po-Lun Ma
Panagiotis Stinis

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
email: orders@ntis.gov <https://www.ntis.gov/about>
Online ordering: http://www.ntis.gov

PNNL-30561

Developing a simulator-based satellite dataset for using machine learning techniques to derive aerosol-cloud-precipitation interactions in models and observations in a consistent framework

September 2020

Po-Lun Ma
Panagiotis Stinis

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

Aerosol-cloud-precipitation interactions (ACPI) remain a major uncertainty in understanding the Earth's radiation budget and water cycle (including extremes). After decades of active research, various observationally based metrics have been developed to constrain ACPI in Earth System Models (ESMs), but direct comparison of model and data estimates can confound scientific understanding because limitations and uncertainties in sampling and retrieval procedures may combine with model deficiencies in process representations of ACPI to obstruct understanding. Furthermore, conventional ACPI metrics often vary from one regime to another, and the ACPI process representation in ESMs is also typically derived based on only a limited area/regime (even though the parameterization applies globally). To bridge the gap between models and data and to correctly describe ACPI in all regimes, we propose to construct a new CALIPSO-CloudSat merged dataset that is produced by the same algorithms used in satellite simulators in ESMs, and to use machine learning techniques to derive new ACPI metrics that can be accurately estimated by satellites and can provide meaningful constraints on cloud microphysical process representations in ESMs. The dataset will include measured and retrieved variables for aerosol, cloud, and precipitation from CALIPSO and CloudSat, and environmental variables from meteorological reanalysis. The data will be used to train a neural network to construct the ACPI metrics as a function of environmental conditions. The new ACPI formula will be used to constrain the ACPI in the Energy Exascale Earth System Model (E3SM), and to augment/reformulate the ACPI process representation in the E3SM to improve the simulation of the evolution of the atmosphere under different environmental conditions.

Objective

The primary objective of this study is to develop a new data-driven formula that describes ACPI, so that simulations of water cycle and atmospheric evolution under different environmental conditions can be improved with higher confidence. The proposed work, compared to conventional approaches, aims to better integrate data and model. The unique contribution includes:

1. The observational data used for deriving the ACPI is constructed using the same algorithms used in simulators in ESMs, minimizing the inconsistencies between models and observations, so that the ACPI can be accurately derived by current satellite technology and can be directly applied to augment/reformulate process representations in ESMs; and
2. The new ACPI formula is derived from samples of all environmental conditions using advanced machine learning techniques instead of from samples within a limited area or regime using simple linear regression, so that the formula can provide a holistic view of ACPI in the Earth system.

Background

Even though the research community has actively worked on ACPI for decades, large uncertainties still remain, with two known challenges: (1) calculations are typically performed inconsistently between models and observations, and (2) applicability of the ACPI metrics is commonly limited to one meteorological regime. The first challenge was recently discussed by Ma et al. (2018) in Nature Communications, showing that the validity of the conventional ACPI metrics can be compromised due to observational limitations. Hence, there is an increasing demand of a new observational dataset which consists of aerosol, cloud, and precipitation fields, and is produced by the same algorithms as those implemented in simulators in ESMs, so that inconsistencies between models and data are minimized and that results from data analysis can be directly applicable to models. The PI Po-Lun Ma has the scientific and technical expertise on this subject, so we can make the unique contribution to the community. The second challenge is related to the fact that conventional ACPI metrics are derived by linear regression methods, while ACPI are nonlinear processes. Furthermore, ACPI process representations in ESMs are often derived from data of limited area or regime, but the parameterizations are applied globally. Studies have suggested that the formulations regulating ACPI vary significantly from one regime to another, indicating that the formula embedded in ESMs is insufficient to describe ACPI in all regimes, and that more factors should be considered when formulating ACPI. Deriving an ACPI formula suitable for all regimes by considering all necessary variables has not been achieved in the past due to the complexity of the problem, and can be benefited from novel machine learning techniques. Our goal is to develop a correct formula rather than an optimal set of parameters/coefficients for a wrong formula. The Co-I Panagiotis Stinis has the expertise and experience with these techniques and will perform the task. Lastly, Ma is familiar with ACPI process representations in E3SM, and can implement the new machine learning based ACPI formula in E3SM effectively. In summary, this project leverages the PNNL strength in ESM, satellite simulator, and machine learning, to improve understanding of ACPI which is an issue of scientific significance.

Scientific Basis and Technical Approach

The proposed study will merge the CALIPSO, CloudSat, and MODIS satellite data with meteorological reanalysis data, use the data to train neural networks, and assess the cloud and

precipitation response to environmental changes under different conditions using the new formulation. Specific steps are described below:

1. Data preparation: We merge the high-resolution CALIPSO, CloudSat, and MODIS data with meteorological reanalysis data. The horizontal resolution of the dataset is 20 km. The dataset includes CALIPSO retrieval of aerosol extinction profile and optical depth, CloudSat precipitation flag, MODIS cloud properties, and meteorological fields. These variables will be used to train neural networks.
2. ACPI derivation: We use *supervised* learning, where we form pairs of tuples (vectors) mapping the environmental variables to the aerosol, cloud, and precipitation variables, and train a generative model, i.e., a neural network, that produces ACPI as a function of environmental tuple. We compare results derived from other machine learning techniques such as random forest and linear regression.
3. E3SM application: We will compare the ACPI derived from the new dataset and from E3SM, and augment/replace the ACPI process representations in the model with the new data-driven formula. Lastly, we will perform short E3SM simulations to assess the evolution of atmosphere with the new ACPI formulation.

Summary of Scientific Results

We have collocated the dataset, combining satellite retrievals with meteorological reanalysis. The data is used to derive cloud response to aerosol perturbations. As shown in Figure 1, we compares the true values of cloud droplet size (as a function of aerosol loading and other environmental variables) to the values of cloud droplet size predicted by DL and by liner regression, finding that the deep learning (DL) produces larger spread of cloud droplet size (when linear regression predicts larger small droplets and smaller large droplets), in better agreement with the truth, but the results are more scattered. A sensitivity test shows that using a deeper neural net reduces the scatter slightly and requires less iteration. By removing one variable at a time, we find that the variable lifting condensation level (LCL) is critical for correctly predicting the spread of the droplet size. When LCL is not included in the training dataset, DL and linear regression produce similar droplet size spread. The results show that using DL provides insights into meteorological factors that affect how aerosols influence clouds even if the effects are non-linear, which cannot be revealed when using linear regression. This sensitivity test also shows that two meteorological variables (estimated inversion strength and humidity) that are commonly used for determining meteorological regimes for understanding ACPI do not actually have a statistically significant impact on ACPI, suggesting that a different way of defining meteorological regimes is necessary for understanding ACPI.

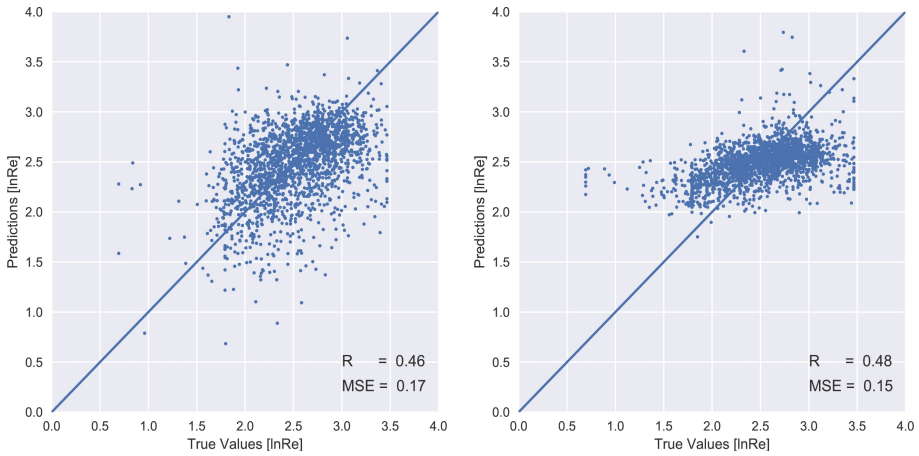

Figure 1. Scatter plots comparing predictions from deep neural network (left) and multiple linear regression (right) with satellite data (true value).

We also used the random forest technique and was able to achieve a remarkable prediction of cloud optical depth probability distribution with predictions bounded by a factor of 2 (not shown). This provides us confidence in using the emulator to explore cloud response to aerosol variations. Using the emulator, we find that cloud optical depth increases with increasing aerosol optical depth (AOD) when the environment is clean (AOD < 0.5), and then the effects saturate so that the cloud optical depth does not change with increasing AOD. A large variability exists which indicates the significant contributions from other environmental (meteorological) variables.

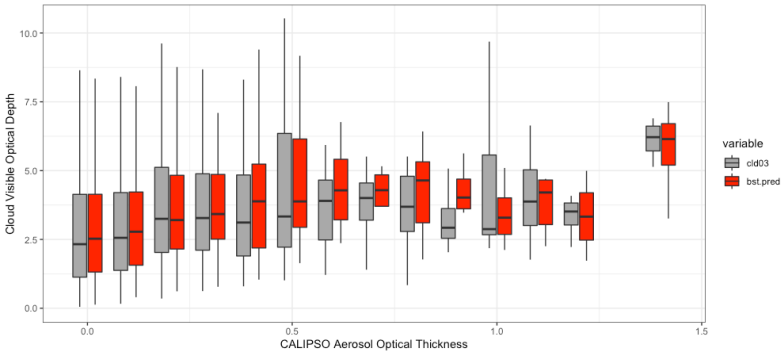

Figure 2. Box plots of cloud optical depth as a function aerosol optical depth derived from satellite (gray) and machine learning emulator (red).

We next analyzed the SHapley Additive explanation (SHAP) values to attribute the variation of clouds to aerosol and meteorological conditions. As shown in Figure 3, the full dataset tries to derive the variation of cloud optical depth as a function of cloud geometric depth (cld14), cloud base height (cld12), surface skin temperature (era4), lower tropospheric stability (era6), precipitation flag (cld15), boundary layer height (era3), relative humidity near surface (era7), column averaged relative humidity (met4), AOD (aer1), cloud effective height (cld10), total column water vapor (era2), cloud top height (cld11), precipitable water (met3), LCL (met1), vertical velocity (era5), and aerosol index (aer2). We use cluster analysis to group clouds into 3 different categories and perform the same analysis. Our results show that the SHAP values and the order of them are very similar to the global SHAP values, indicating that the relationship derived from the emulator is independent of cloud regimes.

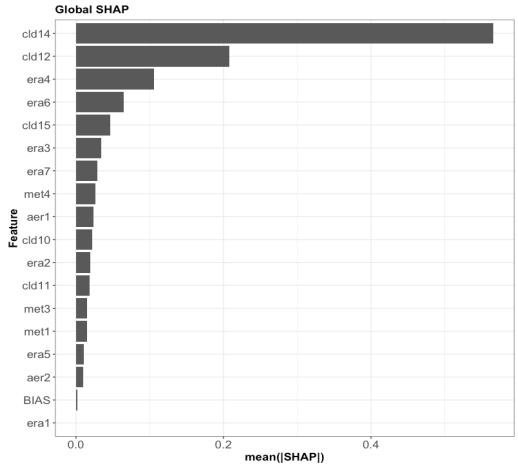

Figure 3. SHAP values ordered by the importance of variables.

Impact

Presentation: This work has been presented at several venues as follows.

Ma, P.-L. (2020), Better cloud calibration leads to improved realism in global atmospheric simulation. Global Model Cloud-Aerosol Research (GM-CAR) Workshop of the 2020 U. S. Climate Model Summit, April 16, 2020, online meeting.

Ma, P.-L., (2019), Using deep learning to derive aerosol-cloud interactions from satellite observations, PNNL TechFest, June 6, 2019, Richland, Washington, USA.

Ma, P.-L., (2019), Using deep learning to derive aerosol-cloud interactions from satellite observations, UCP, February 28, 2019, Berlin, Germany.

Publication: A manuscript is in preparation and will be submitted next year.

Response to funding solicitations: The approach of using novel machine learning techniques to derive observational constraints, and to emulate physical processes, of ACPI have been proposed as part of the "Enabling Aerosol-cloud interactions at GLobal convection-permitting scalES (EAGLES)" project (PI: Po-Lun Ma), funded by U.S. DOE, Office of Science, Office of Biological and Environmental Research (BER), Earth System Model Development (ESMD) program (funding 4M/yr, 2019-2021). Further research on ACPI in E3SM and new observations will be conducted in the EAGLES project.

New Collaborations: Through this LDRD project, we have established collaborations with PCSD scientists including Panagiotis Stinis and Rama Tipireddy. We have also established collaboration with pioneers in using machine learning in Earth system modeling including Christopher Bretherton at University of Washington and Michael Pritchard at University of California, Irvine. Both of them are funded collaborators of the EAGLES project. Furthermore, we established collaboration with Mathieu Vrac at Laboratory of Climate and Environmental Sciences in France.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354
1-888-375-PNNL (7665)

www.pnnl.gov