Geophysical Retrievals in an Artificial Intelligence (AI) Framework for Illuminating Processes Controlling Water Cycle

Virendra P. Ghate, Maria P. Cadeddu, Amanda Lenzi, and Zhengchun Liu
Argonne National Laboratory

Focal Area(s): This white paper responds to Focal Area #3: Insight gleaned from complex data (both observed and simulated) using AI, big data analytics, and other advanced methods, including explainable AI and physics- or knowledge-guided AI.

Science Challenge: This white paper addresses the water-cycle and data-model integration grand challenge. It leverages data from the Atmospheric Radiation Measurement (ARM) Climate Research Facility, and Next-Generation Ecosystem Experiment (NGEE), and Science Focus Area (SFA). The white paper focuses on controlling cloud, precipitation, and radiative properties as observed and simulated by the Earth System Models (ESM). The described framework can be readily applied to any other ensemble of instruments, including satellites and other ground-based networks.

Rationale: The amount of water in the gaseous and liquid phases in the atmosphere (clouds and precipitation) is an essential component of the Global Water Cycle. The horizontal and vertical transport of total water mixing ratio (sum of the vapor and liquid) is solved deterministically in the dynamic core of ESMs through Navier-Stokes equations. However, the formation of clouds and their dissipation through evaporation and precipitation in each model time-step occurs at spatial scales much more refined than the model grid resolution (100 m to 20 km). Hence, clouds need to be parameterized using resolved-scale parameters[1]. This represents one of the most significant uncertainties in the forecast of the future climate and could be addressed by improving our understanding of processes controlling the water vapor-cloud-precipitation transitions at a range of spatial and temporal scales relevant to the ESM. This is further complicated because most of the clouds evaporate with only ~10% of the cloud water converted to precipitation, and some precipitation evaporates before reaching the surface. Some of the processes in this water vapor-cloud-precipitation linkages include i) turbulence that transports vapor from the surface to the condensation level, thereby forming cloud droplets, ii) collision-coalescence process that converts cloud sized drops into precipitation size drops, iii) evaporation of cloud drops due to entrainment of environmental air into the cloud layer and evaporation of precipitation in the sub-saturated sub-cloud layer.

The water vapor, cloud and precipitation properties are not directly measured by any instrumentation but instead need to be derived from remote sensing observations of indirect variables such as sky radiances or backscatter[2,3]. Many novel retrieval algorithms that use data from a single instrument or combine data from multiple instruments have been employed for this purpose. Application of these retrieval algorithms to large amounts of data is very time-consuming due to the required human effort and expertise. It involves instrument calibration, noise-filtering, and re-gridding to a uniform geophysical grid before applying the retrieval algorithms, among other tasks. Moreover, the retrieval algorithms are usually tailored towards particular weather or cloud conditions, hence involving the identification of weather states. For large field campaigns (e.g., MC3E, VOCALS, etc.) that involve multiple instruments, retrievals are produced from data collected by one or more instruments and then re-gridded on a uniform temporal and spatial grid. Such geophysical retrievals are then used for performing process-level studies, evaluating model simulations, and tuning/improving model parameterizations. Although instruments part of the

DOI: https://doi.org/10.2172/1769714
DOE BER, AI4ESP 2021 #AI4ESP1050 (https://www.ai4esp.org)

surface networks and those onboard satellites operate continuously, new insights into atmospheric processes are rarely gained, barring thorough studies that do the tedious work of the tasks mentioned above. The inclusion of all these data is also labor-intensive and computationally intensive due to its large volume. Hence, most of the studies are based on few cases of particular weather/cloud state and rarely utilize more than 10% of the entire collected data (order of Petabytes)[4].

The ARM observatories have been making detailed continuous observations for several decades, with the collected data volume well over several petabytes. Data on heavy precipitation is routinely collected by Weather Surveillance Radar (WSR), 88 Doppler (WSR88D), and several other ground-based networks like Aerosol Robotic Network, Interagency Monitoring of Protected Visual Environments, AmeriFlux, etc., observe other atmospheric state variables. These networks provide detailed information of the atmospheric state at a single point (stations) or over a few kilometers (scanning radars), limiting their application for performing process-level studies and model evaluation. The data from instruments onboard polar-orbiting and geostationary satellites provide crucial top-down information about the aerosol, cloud, water vapor, and radiative properties over the globe. Due to retrieval limitations, aerosol and cloud properties can only be retrieved during the daytime by the polar-orbiting satellites, thereby providing one or two observations per day at a 250m - 3 km spatial scale. Also, the retrieval of aerosol and water vapor properties can only be made during cloud-free conditions. Cloud properties can only be retrieved for clouds that span through the instrument's entire field of view[5].

Narrative: We present an end-to-end AI-driven framework (see Figure 1) that can streamline and expedite data processing and merging data from multiple instruments to yield the state of the atmosphere related to the water vapor-cloud-precipitation interactions at various temporal and spatial scales. Together with advanced statistical techniques, this framework can be used to yield new insights into the workings of the atmospheric processes, thereby paving the way to extract the full scientific value of the collected data. It will also help to identify inaccuracies in the representation of these processes in the ESM. The raw and quality controlled (noise filtered, debiased, calibrated) data from each instrument, along with the retrieved properties of the state of the atmosphere over the last few years, will serve as an input to this framework. The framework has four distinct components, as described below.

(A) First, a series of AI networks (one for each instrument) will transform the raw data into quality-controlled data. The presence of noise is familiar in signal processing, and deep neural networks (e.g., convolutional neural networks (CNN) and generative adversarial networks (GAN)) have long been used as a denoising tool for images as well as one-dimensional signals. Specifically, a generative adversarial network with an encoder-decoder architecture (i.e., denoiser once trained) to map the clean and noisy signal, and another discriminator as a helper to train the denoiser to generate a more realistic clean signal. Hence a CNN for spatial data and Recurrent Neural Network (RNN) for time series or spatiotemporal data can be used to surrogate the entire retrieval process. Specifically, the AI/ML model will learn from imperfect quality-controlled data produced by human experts and the corresponding instrument raw data. The ML model architecture will automatically approximate the relationship between observation data and the target properties mechanism by adequately adjusting the model training regulation mechanism. Thus, once the model is trained, it can be used to process the observational data without any human intervention.

(B) The filtered data from each instrument generated from (A) is used to retrieve properties of the atmospheric state related to the water vapor-cloud-precipitation interactions at various temporal and spatial resolutions. In this step, building confidence intervals for the retrievals is crucial since

each of the instrument's reported parameters has uncertainty. Bayesian Neural Networks (BNN)[6] provide a natural way to quantify uncertainty by assuming distributions over the weights and biases and minimizing the corresponding posterior distribution. A series of AI networks will be needed for generating the entire state of the atmosphere, with one network dedicated for each retrieval technique.

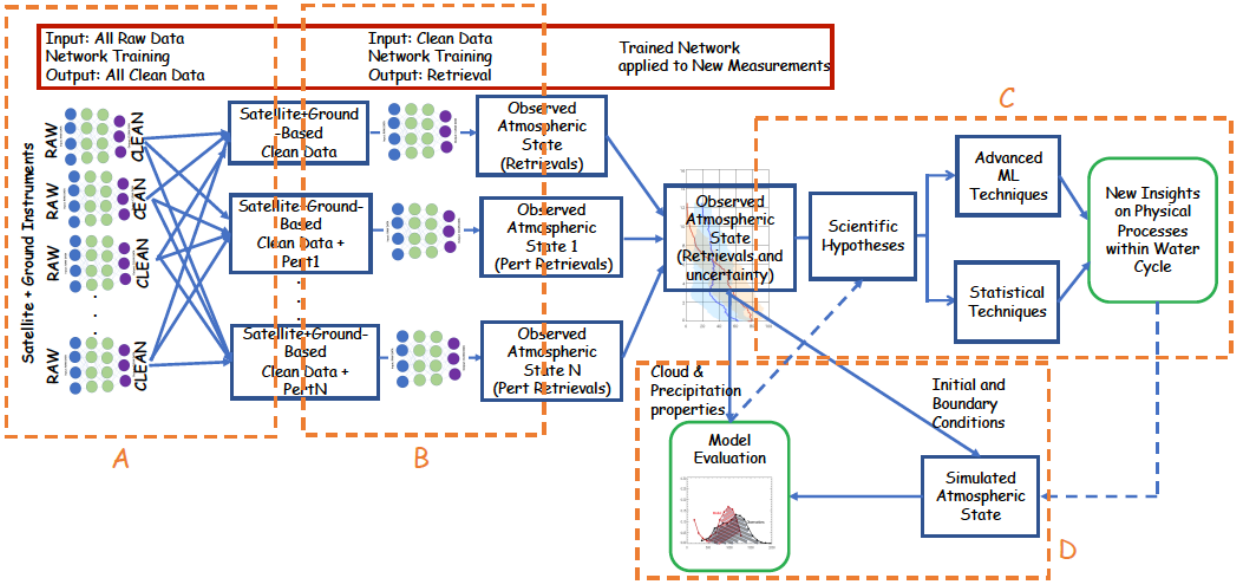

Figure 1: Framework that synergistically uses AI/ML and statistical techniques together with physics based geophysical retrieval techniques to gain new insights on atmospheric processes in water cycle and evaluate models. The four different modules that are alphabetically labelled are shown in orange dashed boxes.

(C) The large amount of data generated in (B) will encompass many atmospheric states. Hence, scientific hypotheses will be critical for yielding new insights into the workings of the water cycle. These hypotheses will be generated through current ESM deficiencies and could be tested using statistical approaches [7]. The first step consists of defining a null and an alternative hypothesis. A z-test (larger data sets) or a t-test (smaller data sets, when the Central Limit Theory breaks) is applied to the data set, which will result in a p-value, which is used to reject the null hypothesis possibly. Meanwhile, advanced ML techniques can provide an insightful explanation for qualitatively evaluating hypotheses. For example, an unsupervised Autoencoder artificial neural network can be trained to extract the most relevant features from high-dimensional observations for visually exploring the similarity and diversity of observations and the water cycle.

(D) The last component of the framework pertains to model evaluation. A few sets of weather conditions (e.g., fair-weather day) could be chosen for performing the model evaluation with the initial and boundary conditions required for performing model simulations being extracted from the generated dataset. Simultaneously, the detailed cloud and precipitation properties may be used to evaluate the model performance. Inaccuracies in the representation of processes within the weather state under consideration can be captured by comparing the simulated atmospheric conditions with those observed. With a sufficient number of these model simulations, metrics such as the Kulback-Leibler divergence[8] could be used to quantify the similarity between the probability distributions of model-simulated and observed parameters.

References

[1] Randall, D. A., A. D. Del Genio, L. J. Donner, W. D. Collins, & S. A. Klein, 2016: The Impact of ARM on Climate Modeling, *Meteorological Monographs*, **57**, 26.1-26.16.

[2] Shupe, M. D., J. M. Comstock, D. D. Turner, & G. G. Mace, 2016: Cloud Property Retrievals in the ARM Program, *Meteorological Monographs*, **57**, 19.1-19.20.

[3] Stith, J. L., D. Baumgardner, J. Haggerty, R. M. Hardesty, W. Lee, D. Lenschow, P. Pilewskie, P. L. Smith, M. Steiner, & H. Vömel, 2018: 100 Years of Progress in Atmospheric Observing Systems, Meteorological Monographs, 59, 2.1-2.55

[4] McCord, R., & J. Voyles, 2016: The ARM Data System and Archive, *Meteorological Monographs*, **57**, 11.1-11.15.

[5] Ackerman, S. A., S. Platnick, P. K. Bhartia, B. Duncan, T. L'Ecuyer, A. Heidinger, G. Skofronick-Jackson, N. Loeb, T. Schmit, & N. Smith, 2019: Satellites See the World's Atmosphere, *Meteorological Monographs*, **59**, 4.1-4.53

[6] Neapolitan, R.E., 2004: *Learning Bayesian networks* (Vol. 38). Upper Saddle River, NJ: Pearson Prentice Hall.

[7] Casella, G. & Berger, R.L., 2021: *Statistical inference.* Cengage Learning.

[8] Anderson, D. & Burnham, K., 2004: *Model selection and multi-model inference.* Second. NY: Springer-Verlag, 63 (2020), p.10.