Quality Data Essential for Modeling Water Cycles Effectively

John Wu, Deb Agarwal, Boris Faybishenko, Marta Gonzalez, Junmin Gu, Tianzhen Hong, Ling Jin, Alina Lazar, Alex Sim, C Anna Spurlock

The experimental and observational data is the backbone of effective modeling of all aspects of water cycles. As the computer technology used for these data gathering efforts has improved dramatically in the recent decades, the environmental conditions these instruments are subjected to remain pretty much the same. These measuring instruments are exposed to the sun light, wind, rain, snow and ice; and they are immersed in dirt, water and mud. All of these are damaging to the electronics components used for data gathering, storage, and transmission, which creates many different kinds of data quality issues that affect the quality and reliability of modeling and prediction efforts.

A critical review of literature indicates that around 80% of a data scientist's time is devoted to finding and organizing quality datasets. Effectively addressing the data quality issues would reduce this data wrangling time and improve the quality of later analysis steps. In this brief note, we describe a number of broad issues related to data qualities, including recognizing anomalies, understanding their impact on the common analysis procedures, and co-designing mitigation procedures with application scientists.

Recognizing anomalies is the first step of addressing the data quality issues. Some anomalies are easy to recognize because an expected data record is missing, say, due to a sensor breaking down or the cellular transmission failing. However, many other situations are not nearly as obvious, for example, a thermometer is reading slightly higher temperature values because the exhaust of the recording equipment is periodically blown toward it by the wind, or the calibration of some measuring devices might be drifting imperceptibly with the seasons. Such subtle anomalies could affect sensitive analysis procedures. Considerable amount of effort is being spent on reprocessing large-scale measurement data to ensure data quality. Much of these efforts are still spent on recognizing subtle anomalies. As additional types of anomalies are discovered, we anticipate further research efforts are necessary to address this issue.

Understanding impact is the next step after recognizing the anomalies. In some use cases, the small perturbations to the data might have negligible impact, while other analysis procedures might amplify the small changes in the input data to produce unexpected results. More mature tools, such as the traditional statistical tools, have well-established properties about their sensitivity and robustness, however, many of the newer analysis tools, such as those complex neural networks, have little or no comprehensive study of their stability characteristics. Uncertainty quantification is being proposed as a possible approach to address this issue, however, it is still an open question whether such an analysis tool could yield convincing results on many layers of neural network with millions of artificial neurons.

Co-designing mitigation is a possible approach that could produce useable strategies to fix the data quality issues in modeling water cycles. For example, the noisy distributed acoustic sensing (DAS) data could be stacked to produce reliable measurements of underground water level because the stacking approach has been validated in many other use cases from geoscience, security, and transportation. The stacking approach was designed by geoscientists who developed the DAS sensing as well as mathematicians who understood the characteristics of the noise. Similar collaboration, i.e., co-design, is necessary to mitigate the subtler anomalies for complex analyses such as modeling the cloud formation for climate simulations. Another important reason for involving both mathematicians and application scientists is that water cycle modeling often involves integration of data from many sources, where the impact of subtle perturbations can be extremely complicated. In such data integrations, each data source has its own particular anomalies to consider, fully understanding the impact of the model would

DOI: https://doi.org/10.2172/1769769
DOE BER, AI4ESP 2021 #AI4ESP1145 (https://www.ai4esp.org)

require input from multiple experts on each individual data source as well as experts on overall data integration procedure.

The co-design process has another critical function in disseminating the math and computer science knowledge to the application domains. In a number of different DOE workshops, trust was brought up as a common concern by application scientists, especially in regards to the analysis tools such as deep neural network. Having application scientists involved in the design of the data quality algorithm would help the application scientists accept the data and the analysis procedures.