Hydrology Data Challenges
Input to the DREAM Team
October 2016

Objective: Provide an integrated data assimilation environment from massive models and observational data sets enabling interactive data analysis and tradeoffs for understanding water availability.

Challenges: The following identify key computing challenges, some of which can be addressed by DREAM, related to establishing capabilities for working with and analyzing hydrology data.

1. **Access to distributed observations**

Observational data are acquired by satellites, airborne and other platforms which *capture, ingest, manage, and archive data within the confines of their systems*. *Access* and use of these data, as part of an assimilation system, are critical to answering questions about water availability. Example observations include: InSAR, Grace, etc.

2. **Access to distributed model output data**

Output from the execution of models are managed within the *confines of their systems making availability and access and important requirement*. These model outputs, like observations, are useful for understanding water availability, particularly for specific types of measurements (river flow, precipitation, etc). Observations can play a key role in helping to generate models. Example: RAPID model, VIC/RHEAS, etc.

3. **Access to InSitu data**

InSitu measurements can be extremely useful but can be more difficult to obtain. They can also have larger *data quality issues* depending on how the data is collected and distributed. *Access* to their data in any *reliable* form may be limited. Examples: well height data.

4. Integration of heterogeneous data and data fusion

Answering broader questions on water availability requires access to a variety of different types of data, as identified above. One of the big challenges is the variety of data which includes significant differences in spatial, temporal measurements, data formats, etc. *Bringing these data together present significant challenges for data fusion.*

5. Orchestrating different model execution environments

Each of the modeling environments perform "model runs" within the confines of their compute environment. In addition to how the data is distributed, as described above, each modeling environment must resource its own compute capabilities to support execution of those models. *Future needs in terms of model execution could be in the areas of integration of modeling and data analysis such that models could be run as part of an analysis environment.*

**6. Optimizing data movement**

Significant challenges existing in the area of data movement. This can affect how a distributed architecture can operate. Decisions over when to move and when to reduce data can have significant impact on the usability of a system. The use of remote compute environments may hinder the effectiveness of the system (e.g., moving data to/from cloud for processing). *Understanding how to optimize the architecture to limit data movement is an important element of scaling.* This is especially true in hydrology with the large models and increasing observational data sets.

7. Working with different archive and data containers

Heterogeneity of data fomats has been raised as an issue, but how data is stored, accessed, and integrated also presents a challenge. Observational data is often captured in "archive containers". To make use of that data, it is often taken from that environment and then put into an environment for running different types of computational analysis. This is particularly true as the shift towards more interactive data analysis is occurring. One of the challenges in enabling interactive data analysis is *putting the data in an environment that is ready to run interactive analysis* on the data.

8. Running analytics-real time

Establishing an environment of interactive data analysis also can lead to the need to run analytics on the data real-time. Rather than running numerous, offline jobs, having data online can be extremely useful for working with the data. This requires having *sufficient computing power and access to the data in appropriate containers that will scale to support data exploration* with reasonable amount of response time.

9. Integrating a visualization platform

Exploration of the data, including the assimilated models and observational data, can be extremely useful, particularly for driving interactive data analysis. Integrating visualization with real-time analytics presents a powerful model for working with massive data, but has *challenges, particularly with rendering massive, heterogeneous data formats and measurements in a meaningful way, as mentioned.*

10. Integration of machine learning and analytics as a service

Establishing a data environment with data in appropriate containers and served by sufficient computing power enables a baseline capability for analytics. Furthermore, the introduction of computational methods to explore and exploit these capabilities (e.g., identifying anomalies in the data) can be used and then expressed through visualization and other means. *Establishing a layer of the architecture that can be used to run analytic methods with the established data ecosystem (scalable computing, access to data, data in appropriate containers, optimized data movement, etc) will enable running computational methods* which can help to automatically detect and/or classify the data.