# ARM Reprocessing Toolkit: Towards Efficient and Timely Delivery of Quality Controlled ARM Data

**Jitendra (Jitu) Kumar, Michael Giansiracusa, Alka Singh, Bhargavi Krishna, James Tonkin, Kavya Guntupally**

Oak Ridge National Laboratory, Oak Ridge, TN

## What's *Reprocessing*?

Data sets in ARM Data Center go through *Reprocessing* whenever a problem is identified in a datastream or an improved processing algorithm is developed to generate a data set. Objective of *Reprocessing* is to ensure delivery of highest quality and accurate data to the scientific community.

## *Reprocessing* Workflow

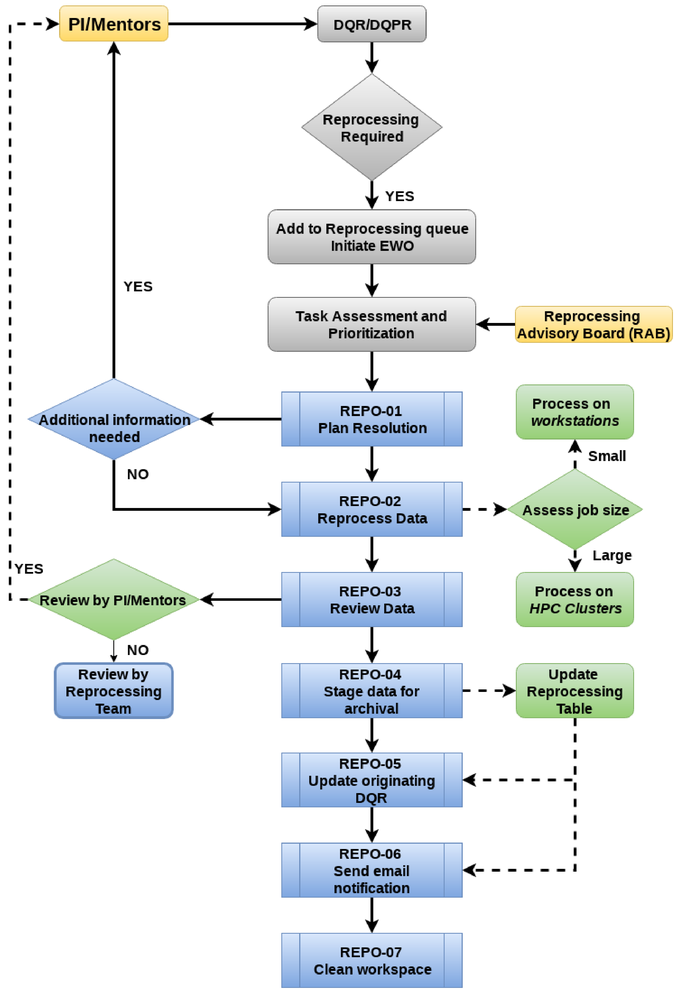


**Figure 1:** *the Reprocessing workflow includes a series of steps to process the data as per DQR, apply versioning, re–archive the corrected data set and notify relevant users/PIs/mentors of the updates to the data.*
*(*Steps followed by current workflow; *Future planned enhancements)*

## Improving How *Reprocessing* is Requested

- New DQR submission form (Figure 2) includes *Reprocessing* specific fields
- Capability to provide symbolic equations that should be applied to correct the data (variable names will be self populated for the selected datastream)
- DQR submitter can request to review the reprocessed data before archival (comparison statistics and plots URL provided to submitter)

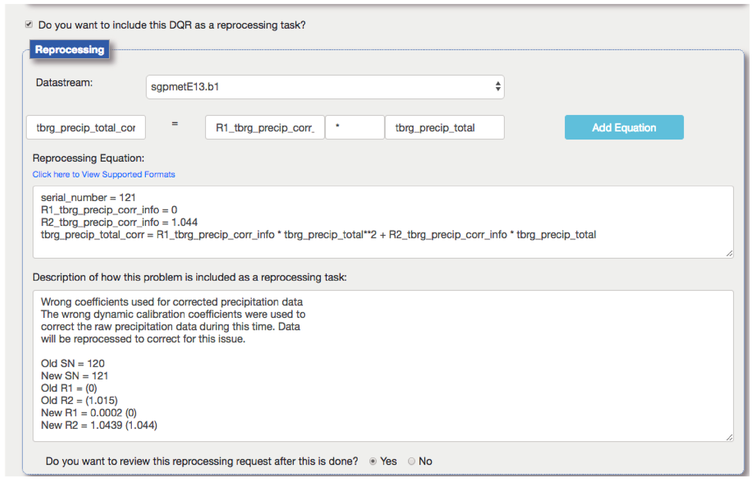


**Figure 2:** *New DQR Submission tool includes options to request reprocessing and provide additional information that the reprocessing team would need to correct the data set (Please see, Guntupally et. al., Poster # 162).*

## Automation for Improved Efficiency

*Reprocessing* often involves a large volume data set which require a series of corrections applied to it. To enable timely resolution of any data quality issues identified by a DQR, we are building a computationally efficient Python-based "**Reprocessing Toolkit**" to automate the workflow.

- **Assess the complexity** of a task and provision jobs on workstations or HPC clusters based on their size and complexity
- **Identify and stage data** (raw and/or NetCDF) necessary for a task using Globus protocols.
- **Process and apply the symbolic equations** to correctly recompute the affected variables
- **Log provenance** information to the database
- **Update DQR** to reflect the modifications
- **Notify users, PIs, and mentors** of the affected datastream via auto-generated email

## Data Dictionaries

Instrument specific data dictionaries allow for accurate and automated processing of the data.

- Encodes the data file formats based on the handbook
- Maps standard variable names
- Includes equations for derived variables in symbolic form
- Dependencies for derived variable
- Captures changes in variable names, instruments, and data formats over time

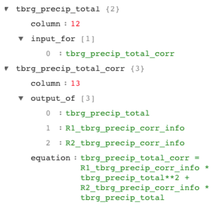


**Figure 3:** *Data dictionary for MET instruments*

## We Welcome Your Feedback

***If you are a PI/Mentor***

- How do we better collect information necessary for a task?
- How to develop priority and time line?
- How do we ensure accuracy of reprocessing (review)?
- How do we improve communication?

***If you are a data user***

- How do we quantify and communicate the effect of a reprocessing task?

Share your ideas, suggestions, wishlists! Leave us a note here or send us an email.

## Contact

Jitendra (Jitu) Kumar, Oak Ridge National Laboratory
Email: kumarj@ornl.gov

## Acknowledgments

The ARM Climate Research Facility is sponsored by the Climate and Environmental Sciences Division (CESD) of the Biological and Environmental Research (BER) Program in the US Department of Energy Office of Science. Oak Ridge National Laboratory (ORNL) is managed by UT-Battelle, LLC, for the US Department of Energy under Contract No. DE-AC05-00OR22725.