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

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)

Oak Ridge National Laboratory, Oak Ridge, TN

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) Do you want to include this DQR as a reprocessing task? Reprocessing Datastream: sgpmetE13.b1 tbrg_precip_total_cor tbrg_pro R1_tbrg_precip_corr_ Reprocessing Equation: Click here to View Supported Formats serial number = 121 R1_tbrg_precip_corr_info = 0 R2_tbrg_precip_corr_info = 1.044 tbrg_precip_total_corr = R1_tbrg_precip_corr_info * tbrg_precip_total**2 + R2_tbrg_ Description of how this problem is included as a reprocessing task: Wrong coefficients used for corrected precipitation data The wrong dynamic calibration coefficients were used to correct the raw precipitation data during this time. Data will be reprocessed to correct for this issue Old SN = 120 New SN = 121 Old R1 = (0) Reprocessing Old R2 = (1.015)Advisory Board (RAB) New R1 = 0.0002 (0) New R2 = 1.0439 (1.044) Do you want to review this reprocessing request after this is done? Yes No Process on workstations Figure 2: New DQR Submission tool includes options to request reprocessing Small and provide additional information that the reprocessing team would need to correct the data set (*Please see, Guntupally et. al., Poster # 162*). Large Process on Automation for Improved Efficiency HPC Clusters *Reprocessing* often involves a large volume data set which require a series of Update corrections applied to it. To enable timely resolution of any data quality issues Table 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, Pls, and mentors of the affected datastream via auto-generated email

\$		
ecip_total	Add Equ	ation
precip_corr_info * tbrg_precip_total		
		/

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

tbrg_precip_total {2

column : 12

instruments

- 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

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?

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

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-AC05-000R22725.



Data Dictionaries

input for [1] 0 : tbrg precip total corr v tbrg_precip_total_corr {3} column : 13 ▼ output_of [3] 0 :tbrg_precip_total 2 : R2_tbrg_precip_corr_info tion : tbrg_precip_total_corr R1_tbrg_precip_corr_info * tbrg_precip_total**2 + R2_tbrg_precip_corr_info * tbrg precip total Figure 3: Data dictionary for MET



Contact

Acknowledgments

