U.S. DEPARTMENT OF ENERGY | Office of Science

DOE/SC-ARM/TR-093

VAP Development: Initiation, Development, Evaluation, and Release

M Jensen
S Collis
J Fast
C Flynn
J Mather
S McFarlane
J Monroe
C Sivaraman
S Xie

February 2011

ARM
CLIMATE RESEARCH FACILITY

DISCLAIMER

DOE/SC-ARM/TR-093

VAP Development: Initiation, Development, Evaluation, and Release

M Jensen
S Collis
J Fast
C Flynn
J Mather
S McFarlane
J Monroe
C Sivaraman
S Xie

February 2011

Work supported by the U.S. Department of Energy,
Office of Science, Office of Biological and Environmental Research

Contents

1.0 Introduction 1
2.0 Stage 0: Initiation 1
3.0 Stage I: Development 2
4.0 Stage II: Evaluation 3
5.0 Stage III: Release 5
6.0 References 7

1.0 Introduction

ARM value-added products (VAPs) provide an important translation between the instrumental measurements and the geophysical quantities needed for scientific analysis, particularly model parameterization and development. The production of VAPs is the responsibility of the ARM infrastructure (translators and developers) with guidance from the science working groups. In 2006, an end-to-end review of the VAP production process by the translator team found that there were several key issues and practices that delayed the formal release of VAPs. In particular, a complete and thorough review of the data to ensure high quality, completeness, and adherence to ARM data standards prior to shipping the data to the Data Archive for permanent storage required significant calendar time. In response to this, a new production stage was introduced, the "evaluation product", which provided for a more timely release of new data products by relaxing requirements on the levels of documentation, descriptions of data quality, and robustness. During the "evaluation product" stage, the scientific community is given the opportunity to analyze and work with the data and provide feedback on the integrity, quality, and usefulness of each product. In practice, evaluation products have served the intended purpose of making new data products available to the larger scientific community through the ARM Data Archive (http://iop.archive.arm.gov); however, the evaluation process has been informal at best, and products have languished in this "evaluation" stage for indefinite and extended periods of time.

This white paper provides a plan to formalize the evaluation of newly developed VAPs and a framework for the development of value-added products through four different stages: Initiation, Development, Evaluation, and Release.

2.0 Stage 0: Initiation

A VAP may be initiated via a Principal Investigator (PI) product that has been particularly well-received, a PI algorithm that provides important new information, a recommendation from the working group/steering committee/science and infrastructure steering committee (SISC), or initiated from a particular need, perhaps in development of another VAP. Priorities on VAP development needs are solicited from the working groups during the annual working group and science team meetings and through each working groups' steering committee in cooperation with the translator team. These priorities are dependent on the needs of the science team and the relevant resources available to the infrastructure.

Once a working group/steering committee or SISC has determined that development of a new VAP should begin, that task is assigned to a translator/developer team. The assignment will depend on the expertise and current workloads of the translators and developers. The translator and developer will work with input from the VAP science sponsor developing the given VAP.

One of the first steps in the initiation stage is the creation of an implementation plan that will include:

- Definition of the translator, developer, and science sponsor for the VAP
- Definition of the input/output variables and or datastreams
- List of sites where the VAP is expected to run

- A basic outline of the algorithm
- An approximate timeline for the development process through submission
- An estimate of the expected level of effort
- Definition of a set of internal tests for the validity of the product
- Definition of three beta users for external testing
- Definition of the conditions for which the VAP is designed to operate
- Identification of need for different data levels (i.e., c1 [run real time], c2 [require post-processing])
- The programming language used for development (This should be an environment in common use within ARM infrastructure unless a compelling reason exists to use a different language).

Once developed, this implementation plan should be reviewed by the translator team before development begins.

The final step of the initiation stage is the submission of an Engineering Change Request (ECR) by the translator or developer. The list of reviewers for the ECR should minimally include the translator, developer, relevant working group chairs, science sponsor, Data Quality Office representative, and beta users. Other reviewers are encouraged where appropriate. The acceptance of the ECR as an Engineering Change Order (ECO) represents the official initiation of the VAP. This ECO should then be the repository for progress updates on development, completion of milestones, reports on changes to the implementation plan, and any other relevant information related to the development of the VAP.

3.0 Stage I: Development

Programming development is begun following the implementation plan. This should include a regular re-evaluation of the implementation plan that is reported on the weekly translator teleconference periodically, but at least once every three months. If there are significant changes required to the original implementation plan, the translator team should discuss alternatives (e.g., help from or transfer to a different developer). Once the programming development is complete, the translator/developer will begin the internal testing that was defined in the implementation plan along with providing an example output file to the infrastructure data product contact for the purpose of an initial Data Object Design (DOD) review. At this time, the inclusion of data quality information is strongly encouraged. While meeting the ARM quality control (qc) standards is not yet required, there are often data quality tests that are implemented in the development and testing process that can be helpful during evaluation and may act as precursors to standardized qc fields. It is highly recommended that the VAP developer use the ARM's Datastream Manager tool (https://engineering.arm.gov/dodmgr/DODMgr.html) to create the fields and qc fields. Finally, when the translator and developer are satisfied with the product, it can be submitted as an Evaluation Product.

The VAP manager will keep a record of the progress following the implementation plan for all VAPs from initiation through release. This record will include changes made to the implementation plan and date of last report (via teleconference). The VAP manager will notify the relevant translator if they are nearing the three-month deadline for a re-evaluation of the implementation plan.

4.0 Stage II: Evaluation

Once the translator/developer/science sponsor determine that the validity and quality of a VAP is sufficient, it may be submitted to be shared with the larger scientific community. This stage aims to strike a balance between quickly making new data products available to the larger scientific community and the application of the full set of ARM data standards important for facilitating the discovery and use of ARM data. To this end the submission requirements include:

- Data product name (must follow ARM naming convention)—For VAPs this name should follow the convention:

  (sss)(nn)(inst)(class)(qq)(Fn).(In).YYYYMMDD.hhmmss.cdf

  – sss is the site identifier (e.g., sgp, twp, nsa)
  – nn is the data integration period in minutes (e.g., 1, 5, 15, 30, 1440)
  – inst is the instrument basename (e.g. mwr, mmcr, mpl)
  – class is the VAP "class" (e.g., aod)
  – qq is a qualifier that distinguishes these data from other data sets produced from the same instrument(s)
  – Fn is the facility designation (e.g., C1, E13, B4)
  – ln is the data level (e.g., a0, a1, b1, c1)
  – The nn, inst, and qq fields are optional and should be considered on a case-by-case basis.
- Data product must be in NetCDF format (exceptions may be made in certain cases).
- Variables must follow ARM naming conventions (https://engineering.arm.gov/task/standards-dod.html).
- When possible, each variable should include a definition of the minimum/maximum and variability over a single time-step that is physically reasonable.
- Provide a list of primary measurement/geophysical parameters for this product.
- Provide a README file describing the algorithm (including references if applicable), data files, and information on potential or known issues with the product.
- Provide a follow-up DOD review of an example output file.

The README file will be the only required source of documentation for the value-added product through stages I-II and therefore must include several important pieces of information:

- Contact information for the translator, developer and science sponsor
- A high-level description of the variables included in the product
- A high-level description of the algorithm used to produce the VAP
- A list of the input datastreams and variables
- A list of the output datastreams and variables
- A list of references (if applicable)
- A sample DOD/netcdf header.

At this time, the "birth of a datastream" (BODS) form is filed, and the above information along with a short name (instrument_class) for the VAP and the README file is provided to the infrastructure data product contact. The primary measurements in combination with the short name of the VAP make it possible for the data to be linked to the measurements and instruments page. The evaluation data will reside in the Data Archive at http://iop.archive.arm.gov and is available to the entire scientific community. The delivery of the data product and accompanying README file to the ARM Data Archive is accomplished through cooperation with the infrastructure data product contact. Upon delivery, a release notice is sent to the ASR science team in order to inform potential users of the new product.

During this stage, the product is further evaluated by a team of at least three beta users, the ARM Data Quality Office, the larger scientific community, and the science sponsor. In order to facilitate the evaluation process, the given VAP should be processed and submitted for a minimum of one year at a single ARM site if the VAP operation is unchanged across sites. If a VAP's algorithm or operating mode differs by site, then a year from each site should be provided (this time constraint may be relaxed for AMF deployments of less than one year's duration, or products that are particular to specific intensive observation periods). Quicklook images of, at a minimum, the VAPs primary variables should also be provided following the ARM naming conventions (ECO#13418). A set of evaluation methods should be clearly defined for each beta user that will include the criteria for acceptance/ rejection of the VAP. These methods may include:

- Comparison to similar products
- Evaluation of long time series
- Use as input to other algorithms
- Other.

The purpose of this evaluation period is to determine any issues with the algorithm including the quality of the datastream, the usability within expected frameworks, and the probable level of interest to the broader community. The beta users, including the ARM Data Quality Office and science sponsor, will evaluate the product over a period of (at least) six months (or other pre-defined time period). Their feedback should be reported to the relevant translator and should include:

- Conditions for which the algorithm does and does not provide useful data
- Identification of specific situations where the algorithm does not perform well
- Recommendation on the readiness of the product for release, including a list of necessary and sufficient improvements needed before release.

At the end of the six-month evaluation period, a decision is made to either;

1. Continue evaluation stage—If issues have been identified during the initial evaluation stage but the beta user team believes the VAP is worthy of further development, the VAP will return to the development stage, addressing the issues and undergo a second evaluation. After the initial six-month evaluation period, additional evaluation should be revisited every three months until a different decision is made.

2. End evaluation phase and continue development towards VAP release—Once the beta user team determines that the VAP is acceptable, continued development will occur towards VAP where efforts are aimed at meeting ARM standards.
3. Terminate further development but release data set—In some cases the evaluation by the beta user team may result in the determination that the product is not worthy of further development but could be useful to a portion of the scientific community and therefore should be released to the ARM Data Archive with no plan for further development. Possible reasons for termination at this stage may include: only applicable to a small subset of observations, low impact towards achievement of programmatic goals, lack of users, existence of superior data sets, etc. The accompanying README file should be updated to reflect the termination and underlying reasons.
4. Terminate further development and withdraw VAP—In some cases the evaluation may lead to the conclusion that any further development should be terminated and the VAP should be withdrawn. Possible reasons for this decision may include: identification of scientific errors in algorithm, production of physically unrealistic results, problems with input datastreams, etc. Users who have downloaded the product during its evaluation stage should be notified via e-mail by the infrastructure data product contact.

The status/version of a product in the evaluation stage will be captured in the global variable "process_version", which should be set to "beta_version_<x.x>" where x.x represents the version major.minor number. The version number should be incremented whenever the evaluation product is reprocessed and stored in the evaluation area. The infrastructure data product contact will monitor the data and will make sure that the version is incremented. The change log in the README file should reflect the reason for the update. When/if this data moves to production, the version number will not be updated unless the package is run again on the released code.

5.0 Stage III: Release

If the decision is made to terminate further development but release the VAP in its current form, the BODS form and README file are updated to reflect this decision, including a short description of the justification. The release filename is designated (designation to be determined) to identify the differing status of this type of VAP.

If the decision is made to move towards VAP release, further development is concentrated on improving the accessibility and usability of the data product. This includes improvements in the description of the quality of the data, the documentation, and metadata. In considering the full implementation of ARM standards, there may be cases where meeting these standards is beyond the means of the resources available, or inappropriate for a given VAP. In these cases there is an option of releasing the data set that does not meet all standards with an appropriate filename designation (to be determined) that identifies this status.

During this stage, ARM QC standards are implemented as described in Gaustad et al. (2010). Structuring the VAP output files to meet the ARM QC standards moves the VAP closer to meeting ARM data ingest requirements and ensures the VAP is compatible with existing ARM data quality assessment tools. At the most general level, the QC fields in a VAP data file should:

- Immediately follow the fields for which they report quality
- Use a naming convention of "qc_<fieldname>"
- Be of type integer
- Store the quality information using a bit-packing technique to allow multiple test results to be presented in a single value.

In some cases a VAP translator may want to supply QC information in a non bit-packed format. For these instances, an auxiliary QC field may be created. In addition to referencing the standards documentation, the VAP developer will be greatly aided by ARM's Datastream Manager tool (https://engineering.arm.gov/dodmgr/DODMgr.html), which auto-detects and highlights fields in the DOD that do not support the current standards. Once these standards are implemented, the VAP data file should meet all ARM naming and QC standards. For release, the global variable "process_version" should be changed from "beta_version_<x.x>" to "<x.x>" where x.x represents the version major.minor number. This version number will not be updated unless the package is run again on the released code.

The translator/developer will write a technical report and web page that will include information for users to better understand the algorithms used to process the data product, the variables included in the product, and the quality checks implemented. Before the VAP is released, the beta users and other translators are given the opportunity to provide feedback on the final files and technical document. With approval from the beta users and the translator team, the VAP is ready for final release.

At this time the BODS information is updated with the primary measurements and category. This information is propagated through the ARM web pages, including listing this new VAP on the VAP web page and as a source on the appropriate primary measurements pages, linking the technical report, updating the VAP status report, and ensuring that the Data Archive browser shows the data.

The release process includes compiling code, testing the data to validate the data, providing information about how and where the VAP should be run at the Data Management Facility (DMF), and the required datastreams for the VAP.

The steps included in the release process are:

1. Compile the code under a VAP component.
2. Create a test script that runs the VAP, produces data, and compares with a referential data set to ensure that the data produced is the same. Often a developer links certain libraries in their home environment that are not accessible to the DMF. This process ensures that the VAP that is run in the developer's home environment is the same as it is at the DMF.
3. Create a nconf file that provides package details (i.e., files to be included in the package), technical details for processing the VAP at the DMF (i.e., required datastreams) optional datastreams, sites to

be run, and contact information.
http://engineering.arm.gov/base/swawt/eg.html?template_name=file_pnc&role=developer

4. Generate an RPM package using the genrelease command. This process checks out the code from the repository and runs the test script. Once the test passes, it creates a package for release.
5. A Baseline Change Request (BCR) is written to deploy the package and turn over operations to the DMF. After the BCR is approved, the package is then released by the Configuration Manager to the DMF.

At this time the new VAP datastream is expanded through the processing of historical data that will be included in the ARM Data Archive. The extent of this historical processing depends on the available resources, the user community needs, and the applicability based on availability of input datastreams. Upon delivery of the data to the Data Archive, a release notice is sent to the ASR science team in order to inform potential users of the new finalized product.

Although the preferred mode of operation is for VAPs to be processed by the DMF, in some cases the VAP algorithm may preclude processing by the DMF (e.g., expert interaction is required). If this is the case, upon agreement by the translator team, developer and ARM Data Archive, final approved data products may be transferred directly to the Data Archive through direct arrangements with the ARM Archive Manager. In this situation, the offsite products and underlying reasons should be clearly stated in the VAP technical report.

6.0 References

Gaustad, KL, CJ Flynn, SJ Beus, and BD Ermold. 2010. "The development of QC standards for ARM data products." In *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, Chicago, Illinois, USA.

www.arm.gov

U.S. DEPARTMENT OF
ENERGY

Office of Science