

CERES Systems Engineering Committee

Members: Tammy Ayers, SSAI (CM)
Denise Cooper, SSAI (DM)
Tonya Davenport, SSAI (ASDC)
Thomas Grepiotis, SSAI (DM)
Nelson Hillyer, SSAI (DM/Systems)
Vertley Hopson, SSAI (ASDC)
Victor Sothcott, SSAI (DM)

Charter: Serve as a forum for resolving issues which affect more than one working group. Report to the CERES Data Management Team.

August 2, 2013 10:00 am

The following members were present for this meeting: Denise Cooper, Tonya Davenport, Thomas Grepiotis, Vertley Hopson, and Victor Sothcott.

Walt Miller started the meeting with opening remarks on Configuration Code (CC) values and why we need to document the current scheme being used to set these values. Science Team members are often confused about how these are being set and the exact nature of how and when they are updated. In the past, the 6-digit CCs were set based on internal and external changes that affected a PGE. The first three digits tracked internal changes to the science software. The last three digits tracked external changes that impacted a PGE such as CERESlib updates, operating system changes, and a new Toolkit. Some of these factors have been found to be less important than originally thought and so a new set of rules was developed during the initial phase of Edition3 data products. Unfortunately, these rules were never fully documented and some of the initial Edition3 data products did not follow the new CC scheme, while some Edition1 data products ended up with Edition3 CC numbers.

Lisa Coleman gave a presentation on the strategy for setting CC numbers from Edition3 forward. Under the new strategy the first three-digit number in a CC tracks changes to input data and the second three digits tracks software changes for a PGE. A copy of the slides is included in these notes. Lisa also presented a draft flowchart of the process for updating CC values along with a draft outline of the proposed document to record the new strategy.

The group reviewed the draft flowchart and made suggestions on possible changes. A suggestion was made to break the flowchart into processing streams, and Lisa will work with Joanne Saunders to implement this suggestion. A high-level flowchart for the Science Team and other users will be developed after finalizing the individual processing stream flowcharts. Lisa will provide the updated flowcharts to the SEC for comments and suggestions when they are complete.

Following the presentation there were a number of questions and suggestions. There will be a general rule that CC values will only change when (1) the input data format/files change and/or (2) a software delivery requiring a recompilation occurs. It was proposed that all data products in a processing stream have the same CC value. For example, Instrument and ERBE-like data products would have the same CC values throughout. An exception to the rule is that if for some reason an ERBE-like PGE had to be updated in mid-stream during processing the new CC value would only affect data products downstream of that PGE. Lisa proposed that CC values be set according to Edition numbers when applicable. For example, Edition1 data products would have a CC value of 1xx1yy, where xx is the version of the input data (00-99) and yy is the version of the software (00-99). Also, for PGEs that can run multiple Editions, yy will be the same for every Edition run with the same version of software. As an example, PGE CER1.4P3 can be used to run Edition1, 2, 3, or 4, so any Edition1 products will have a CC value of 100103, Edition2 – 200203, Edition3 – 300303, and Edition4 – 400403. Whenever the software for this PGE is redelivered and recompiled, the last two digits in each of the Edition CC values will be incremented by at least one.

The meeting was adjourned at 10:55 AM.

CERES CC Numbers Edition3. And Beyond.

August 2, 2013

Lisa Coleman

SEC Presentation

Changes from Edition2 Strategy

- Reset values to 300300 at the beginning of processing (for most PGEs – Instrument and ERBE-like began Edition3 processing before this could be implemented)
- Made the 3 leftmost digits indicators of external data source, and the 3 rightmost digits indicators of CERES source code modifications
- No bumps for delta script-only deliveries – so far we have not encountered an external input data change that only required a script change
- Attempting a “same CC number for all products in a family” approach
- Dataset-based in addition to being PGE-based, i.e., if a PGE generates products for multiple datasets, each dataset could have a different CC number

First 3 Digits – Edition3 Publicly Available Products

- The first, or leftmost, three digits indicate something about the external data source
- External input data source consistent (so far) throughout data record for Instrument and ERBE-like products
- Example from another project – GMAO has the 100 series, the 200 series, etc., to distinguish between output resulting from reprocessing and forward processing
- Changes to external input sources for Edition3A SSF-based CERES products (so far):
 - MODIS Collections for Clouds properties
 - GMAO 4 to GMAO 5
 - Edition2 IES to Edition1-CV IES

Edition3A SSF External Inputs

CC 1 st 3 Digits	Dates	MODIS Collection	GMAO	IES
300	Launch – April 2006 (Terra overlap with 301 thru July 2006)	4	4	Edition2
301	May 2006 – November 2007	5	4	Edition2
302	December 2007 – June 2010	5	5	Edition2
303	July 2010 forward	5	5	Edition1-CV
304 ??	MODIS Collection 6 ??? No discussion to date on CERES Edition3 and use of MODIS Collection6			

First 3 Digits – Other Products

Follow same logic

- Clouds Snow and Ice and Edition1-CV SSF
 - Once on AMI, set CC to 200200. At the same time, the external snow and ice inputs also changed.
 - Impacted Clouds Ed1-CV, so the output CC bumped to 201200, and likewise for Inversion Ed1-CV products
- ISCCP-D2like
 - For SSF based products, be consistent with SSF input
 - For GGEO-based products – Running Edition3A for first time, so CC=300300 for entire data record. So far.

Second 3 Digits

- The second, or rightmost, three digits indicate an internal CERES software (source code) change
- Has not worked out to have same CC number for all CERES products for a given date
- Approach of trying to have “product suites” with the same CC number
 - Ed3A SSF, SFC, SAH (Cloud input, single satellite – each of which could have different CC numbers)
 - TSI, SYNI, SYN1deg suite (multi-satellite)
 - Now – IES, BDS, ES8, ES4, and ES9 (no Cloud input)
 - Edition1-CV Clouds and Inversion

Out of Sync CC Numbers Within a Suite

- Typically, the last three digits of CC numbers are incremented by 1 when it is time to increment
- HOWEVER - Edition3A Terra SSFs ended up with different numbers than the Aqua SSFs for some of the same months. At some point it became reasonable to increment the CC by 1 for one satellite/product, and by 2 for the other so that they were again the same
- When an emergency reprocessing that begins with a product other than the first product in a suite is required, the CC numbers within a product suite will get out of sync (e.g., the SYN1 for 5/12 forward). At the next reasonable opportunity, assign consistent CC numbers for all products in suite – may need to increment the CC number for some products by more than 1 in such a case.
- Usually do not want to reprocess products just for the sake of synchronizing the CC numbers

Impacts of Changing CC Numbers

- Necessary and accepted, but not without some impacts
- Requires new PRs to be written for the PGEs producing a product, as well as the PGEs using that product as input
- Impacts SIT-developed scripts for same PGEs
- Writing new PRs means reviewing new PRs
- More information to keep track of

Directives

- Near-term: Outline of a document - internal
- Flow Chart - Public

First (leftmost) 3 Digits

```
IF (output product Edition3) THEN
  Initial value = 300
  IF (output product SSF or SSF-based) THEN
    IF (Cloud inputs MODIS C4-based) THEN
      First three digits = 300
    ELSE IF ( Cloud inputs MODIS C5 based and MOA data are GMAO 4 based)
THEN
      First three digits = 301
    ELSE IF ( Cloud inputs MODIS C5 based and MOA data are GMAO 5 based
and
      SCC Edition2 SSF based) THEN
      First three digits = 302
    ELSE IF ( Cloud inputs MODIS C5 based and MOA data are GMAO 5 based
and
      SCC Edition1-CV SSF based) THEN
      First three digits = 303
    ELSE IF (New change to an external input source occurs) THEN
      First three digits = 303 + Number of changes since use of value 303
    END IF
  ELSE
    IF (no changes in external input source )
      First three digits = 300
    ELSE
      First three digits = 300 + number changes to external inputs since Edition3
processing began
    END IF
  END IF
END IF

ELSE IF (output product (Edition1-CV .or. Baseline1-QC) .and. (Instrument or
ERBE-like) ) THEN
  First three digits = 300
ELSE IF (output product Edition1-CV and (Clouds or Inversion)) THEN
  First three digits = 200 + number changes to external inputs since Edition1-CV
processing began
ELSE IF (output product = Edition 4) THEN
  First three digits = 400 + number changes to external inputs since Edition4
processing began
ELSE IF (output product = Edition 1) THEN
  First three digits = 100
END IF
```

Last (rightmost) 3 Digits

```

IF (output product Edition3) THEN
    Initial value = 300
ELSE IF (output product (Edition1-CV .or. Baseline1-QC) .and. (Instrument or
ERBE-like) ) THEN
    *** Same PGEs used for Edition1-CV, Baseline1-QC, and Edition3 processing
    Initial value = 300
ELSE IF (output product Edition1-CV and (Clouds or Inversion)) THEN
    *** Separate PGEs used for Edition1-CV processing than for Edition3 processing
    Initial value = 200
ELSE IF (output product = Edition 4) THEN
    *** Separate PGEs used for Edition4 processing than for all other processing
    Initial value = 400
ELSE IF (output product = Edition 1) THEN
    *** Same PGEs used for NPP Edition1 processing as for other data sets - maintain
correlation
    with other Edition data ssets using same version of software
    Initial value = 100 + last 2 digits of CC number associated with PGE
END IF

IF (PGE source code change) THEN
    *** Time to increase the CC number
    IF (current value >= values for other PGEs/products in family) THEN
        Increment output CC value by +1
        IF (affected product is input to another PGE in same family) THEN
            Set output CC value of target PGE equal to modified PGE
        END IF
        IF (affected product is input to another PGE in different family) THEN
            IF (no special exception made by DMT lead) THEN
                No change to output CC for product in different family
            END IF
        END IF
    ELSE IF (current value < values for other PGEs/products in family) THEN
        Set new value == greatest value currently used in PGE/product family
    END IF
END IF

```

I. Introduction

- A. What is a CERES CC number?
- B. Why CERES uses CC numbers
- C. CC Numbers under scope of this document
- D. Definitions
 - 1. External Input Product
 - 2. Internal Input Product
 - 3. Production Request
 - 4. Data Set/ Output Production Strategy
 - 5. Product suites

II. Strategy of Assignment of CC Numbers

- A. Initial Values
 - 1. Represent highest Edition quality data set produced by source code
- B. Value Changes
 - 1. First three digits
 - a. External input change in PGE, or in a parent PGE
 - b. Typical amount of increment
 - 2. Last three digits
 - a. PGE Executable change, or change in an executable for a PGE in same product suite
 - b. Typical amount of increment

III. Looking up CC Numbers

- A. DPO file listings
- B. CC Look Up Tool
- C. PRDB Search Capabilities – Also provides an SCCR number for the executable

(The following is from an email written by Jonathan and may be useful)

Why do we need a cc code?

The Configuration Code acts as a unique identifier for the software, hardware, and with Edition 3, input data used to produce a data product. The cc code enables us to identify how the data were produced. Starting with Edition 3, we changed what the cc code is used for.

Edition 2 used letters (e.g. Edition2A, Edition2B, Edition2C) to delineate a change in input data type for a given data product. In this case the cc code only referenced the software and hardware configuration. So for example, when we transitioned from GEOS4 to the G5 GMAO input data stream (used to produce MOA) we have December 2007 as an overlap. In the SSF, for Terra-FM1, December 2007 there is an Edition2F and an Edition2G, both with the same cc code of 028035. Since we used the lettering system, we did not have to increment the cc code for this instance as there was no change in the software or hardware configuration. Likewise, the DPO directory structure is organized as Edition2F and Edition2G etc.

For Edition 3, we eliminated the letters and are using the cc code to account for input changes also. As a general rule of thumb, wherever there is overlap, the higher cc code should be the most recent input/software/hardware configuration. Since we changed this approach for Edition 3, we documented our strategy in the following document

(http://ceres.larc.nasa.gov/documents/dmt_mthly_ceres/pdf/Edition3ProcessingPlan_09-08-2011.pdf). Table 4 references the cc codes.

Note, since the data is no longer organized into directories such as

Edition2F, Edition2G, etc, all overlapping data are in a single directory (such as Terra-FM1_MODIS_Edition3) and we have to use the cc code to identify which data to use.

Why can't we only have the "most recent" data in the DPO?

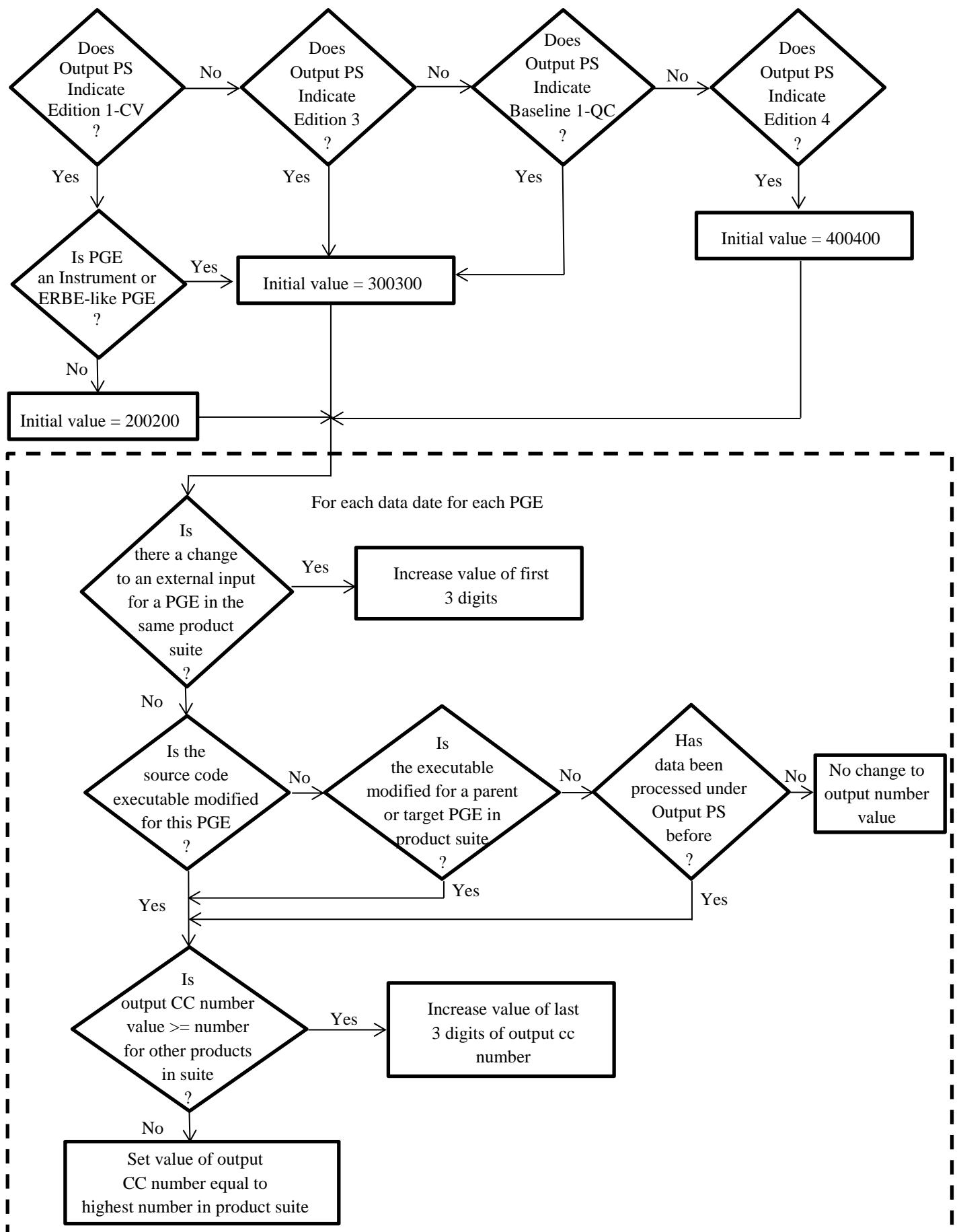
The DPO was designed to house all data that feeds CERES production and offline processing such that data doesn't have to be staged to disk from tape. On AMI, this significantly increased our production capacity. As long as there is a need to process overlap months for offline comparison, these files should be in the DPO.

Why Do we need a cc code in the file name? Why can't we just put it in the metadata?

For production, we specifically call out which input data we want to use for a given production run by listing the cc codes. Lisa Coleman and Jeremy Lande are responsible to do this with the PR. This ensures that the data is produced correctly. The PCF script that runs before each PGE ingests input parameters provided in the PR, including cc code, for all input data such that it can build a list of all input files for the PGE to run. The PCF reads file names, similar to an ls command, to get it's information such that it doesn't have to open all the metadata files to find the correct files. If the cc code were moved out of the file name and into the meta data, we would have to functionally change how the PCF scripts operate. In this scenario, the PCF scripts would have to read all possible metadata files for a given date range into memory to open and examine the cc codes. This would introduce a time constraint and an IO constraint. Furthermore, we would have to rewrite all our PCF scripts, which is doable, but a non trivial effort.

There is a possible light at the end of the tunnel. The document I referenced above has all the information necessary to cross reference with PRs and identify what input is used for a given cc code. We made a table for ASDC that lists all CC codes for Edition 3 data and identifies which should be publically released. The table identifies the "preferable" cc code in cases where there are overlap such that we

only distribute one version to the public. We can make this available. We could build something like this for Edition 2 and earlier as a reference.



Flow Diagram of CERES Output Configuration Code Number Assignment Strategy