|  |  |  |
| --- | --- | --- |
| WORLD METEOROLOGICAL ORGANIZATIONCOMMISSION FOR BASIC SYSTEMS-----------------------------FOURTH MEETING OF INTER-PROGRAMME EXPERT TEAM ONDATA REPRESENTATION MAINTENANCE AND MONITORINGGENEVA, SWITZERLAND, 30 MAY - 3 JUNE 2016 |  | IPET-DRMM-IV / Doc. 4.2.1 (1)(13. 5. 2016)-------------------------ITEM 4.2ENGLISH ONLY |

**Validation report of GRIB2 new compression method – CCSDS libaec**

**(2013-2.2.1 DRMM-I, IPET-DRMM-IV Doc 2.2(6))**

Luis Kornblueh (MPI-M), Mathis Rosenhauer (DKRZ), Enrico Fucille (ECMWF),

Gregor Schnee (DWD), Sibylle Krebber (DWD), Daniel Lee (DWD)

# 1. Responsible organization

Max-Planck-Institute for Meteorology (MPI-M), German Climate Computing Centre (DKRZ), European Centre for Medium-Range Weather Forecasting (ECMWF), German Weather Service (DWD)

# 2. Requirements and purposes

Improved packaging of GRIB data is already available using JPEG 2000 and PNG. We would like to add another algorithm – the standard for lossless data compression recommended by the Consultative Committee for Space Data Systems (CCSDS).

Tests have shown that this compression method is very fast (by one order of magnitude compared to currently available algorithms) and that it reaches a very high compression ratio (a mean factor of 2.7 – 3.0 compared to the standard GRIB encoded records). Several GRIB archives exist on the petabyte scale. Their exponential growth rates make the use of fast compression essential for operational data access, archival and retrieval within data centers. Furthermore, compression reduces transmission bandwidth.

Since the original submission of the proposal to ET-DRC 2008 several issues have been resolved:

* A US patent has been issued to NASA on one algorithm. Following a conversation with NASA's Chief Patent Counsel they have effectively abandoned this patent and anyone is free to make 'use of its teachings and claims' (see documentation in [3]).
* In addition to the non-free previous implementation of the CCSDS recommendation, szip, a free implementation of the standard is now available in the form of libaec.

libaec is a free implementation by a joint effort of DKRZ and MPI-M of the Golomb-Rice compression method as described in the Space Data System Standard 121.0-B-2. libaec has an szip API compatibility layer and can be used as a drop-in replacement for szip. The code is written in ANSI-C and available under a BSD open source license. All szip related tests in HDF5 pass with libaec and the code has been extensively tested with HDF5 and the proposed GRIB extension. The library can be found at [3].

The German Climate Computing Centre, DKRZ, is offering long-term, best effort support.

The proposal is designed based on the previous compression add-ons in data representation templates 5.40 and 5.41.

# 3. Description of proposal

During the validation process the length of the uncompressed GRIB message and the size of the uncompressed data were removed from the template.

A detailed description of the lossless data compression can be found in [1].

**Templates:**

Preliminary note: For most templates, details of the packing process are described in regulation 92.9.4

|  |
| --- |
| **Data Representation Template 5.42 - Grid point and spectral data - CCSDS recommended lossless compression.** |
| Octet No. | Contents |
| 12 – 15 | Reference value (R) (IEEE 32-bit floating-point value) |
| 16 – 17 | Binary scale factor (E) |
| 18 – 19 | Decimal scale factor (D) |
| 20 | Number of bits required to hold the resulting scaled and referenced data values (see Note 1) |
| 21 | Type of original field values (see Code Table 5.1) |
| 22 | compression scheme version number of CCSDS 121.0-B recommended standard blue book (currently 2) (see Note 3) |
| 23 | Block size |
| 24-25 | Reference sample interval (see Note 3) |

Notes:

1. The intent of this template is to scale the grid point data to obtain the desired precision, if appropriate, and then subtract the reference value from the scaled field as is done using Data Representation Template 5.0. After this, the resulting grid point field can be treated as a grayscale image and encoded into the CCSDS recommended standard for lossless data compression code stream format. To unpack the data field, the CCSDS recommended standard for lossless data compression code stream is decoded back into an image, and the original field is obtained from the image data as described in regulation 92.9.4 Note (4).
2. The Consultative Committee for Space Data Systems (CCSDS) recommended standard for lossless data compression is the standard used by space agencies for the compression of scientific data transmitted from satellites and other space instruments. CCSDS recommended standard for lossless data compression is a very fast predictive compression algorithm based on the extended-Rice algorithm. It uses Golomb-Rice codes for entropy coding. The sequence of prediction errors is divided into blocks. Each block is compressed using a two-pass algorithm. In the first pass the best coding method for the whole block is determined. In the second pass, the output of the marker of the selected coding method is encoded as ancillary information along with prediction errors.

The coding methods include:
	* Golomb-Rice codes of a chosen rank
	* Unary code for transformed pairs of prediction errors
	* Fixed-length natural binary code if the block is found to be incompressible
	* Signaling to the decoder empty block if all prediction errors are zeroes
3. Consultative Committee for Space Data Systems: Lossless Data Compression.
CCSDS Recommendation for Space Data System Standards,
CCSDS 121.0-B-2, Blue Book, May 2012.

|  |
| --- |
|  **Data Template 7.42 - Grid point and spectral data - CCSDS recommended lossless compression.** |
| Octet No. | Contents |
| 6 - nn | CCSDS recommended standard for lossless data compression code stream |

**New entry in Code table 5.0 – Data representation template number**

|  |  |
| --- | --- |
| Octet No. | Contents |
| 42 | Grid point and spectral data - CCSDS recommended lossless compression. |

**ANNEX:**

1. Consultative Committee for Space Data Systems: Lossless Data Compression. CCSDS Recommendation for Space Data System Standards, CCSDS 121.0-B-2, Blue Book, May 2012.
2. Consultative Committee for Space Data Systems: Lossless Data Compression. CCSDS Report Concerning Space Data System Standards, CCSDS 120.0-G-2, Green Book, May 2012.
3. Library for adaptive entropy coding from Deutsches Klima Rechenzentrum (DKRZ) - libaec. <https://gitlab.dkrz.de/k202009/libaec>.
4. Reference implementation of CCSDS encoding using wgrib2. <https://github.com/erget/wgrib2>.
5. Patch to ecCodes for bug free CCSDS encoding (will be included in a future release of ecCodes). <https://gist.github.com/erget/9664a4307f1ed9092675faee47cbd6df>.
6. Technical validation proofs, including test data, demonstrating error-free CCSDS encoding and decoding on > 10,000 GRIB messages using the modified implementations of wgrib2 and ecCodes. <https://gist.github.com/erget/0533fd25d83a61af343353e9114e7e7b>

# 4. Declaration of validation complete

Validation was carried out using a heterogeneous collection of 10188 GRIBs. They were collected from several centers and originating processes. The tests verified that both ecCodes and wgrib2 were capable of encoding and decoding data packed with the proposed template without deviating from the original data. This was done by rewriting the input GRIBs using the proposed template, and then in a second process rewriting the GRIBs from the proposed template using simple packing (data representation template 5.0). The test passed if values in both the original and the repacked GRIB using simple packing were identical.

One issue remains in ecCodes: Data encoded using data representation template 5.3 cannot be directly rewritten into the proposed template. This is currently under investigation and will be removed from a future release. In the meantime a workaround exists: If an intermediate data representation template is used, the data can be correctly rewritten using the proposed template (e.g. template 5.3 🡪 template 5.0 🡪 template 5.42).

For technical details, see [6].

# 5. Period of discussion and its conclusion

During validation the proposed template was changed to its current form (the original proposed templates can be found in section 7).

Changes were made to libaec in order to fix a bug that occurred when encoding data with a length that could not be evenly divided into the reference sample interval (RSI, compression input/output bits per pixel und scan line). This can be found in commit e047ec682044d8a25b7111699ec7d977ea065efa.

Changes were made to ecCodes in order to provide support for this template and, subsequently, to enable use with data with 16 < x < 25 bits per value.

Changes were made to wgrib2 in order to provide support for this template.

All changes have been reported back to the maintainers and developers of the respective libraries. They can be found in the annex. Pending approval and integration, they will be included in future releases of the libraries. Thus annex documents 4 and 5 will become obsolete, but their URIs shall remain persistently available for reference.

# 6. Proposed implementation date and its formality

If agreed, November 2016.

# Brief summary of discussion

Minor changes were made to the original proposal to include exactly the amount of information required in the template in order to encode and decode data properly. The previously proposed table changes are shown below:

|  |
| --- |
| **Data Representation Template 5.42 - Grid point and spectral data - CCSDS recommended lossless compression.** |
| Octet No. | Contents |
| 12 – 15 | Reference value (R) (IEEE 32-bit floating-point value) |
| 16 – 17 | Binary scale factor (E) |
| 18 – 19 | Decimal scale factor (D) |
| 20 | Number of bits required to hold the resulting scaled and referenced data values (see Note 1) |
| 21 | Type of original field values (see Code Table 5.1) |
| 22 | compression scheme version number of CCSDS 121.0-B recommended standard blue book (currently 2) (see Note 3) |
| 23 | compression options mask (see Note 3) |
| 24 | compression input/output bits per pixel (see Note 3) |
| 25 – 26 | compression input/output pixels per block (see Note 3) |
| 27 – 28 | compression input/output pixels per scan line (see Note 3)  |
| 29 – 36 | length of the uncompressed grib message in octets |
| 37 – 40 | size of uncompressed data in octets |

Notes:

1. The intent of this template is to scale the grid point data to obtain the desired precision, if appropriate, and then subtract out reference value from the scaled field as is done using Data Representation Template 5.0. After this, the resulting grid point field can be treated as a grayscale image and is then encoded into the CCSDS recommended standard for lossless data compression code stream format. To unpack the data field, the CCSDS recommended standard for lossless data compression code stream is decoded back into an image, and the original field is obtained from the image data as described in regulation 92.9.4 Note (4).
2. The Consultative Committee for Space Data Systems (CCSDS) recommended standard for lossless data compression is the standard used by space agencies for the compression of scientific data transmitted from satellites and other space instruments. CCSDS recommended standard for lossless data compression is a very fast predictive compression algorithm based on the extended-Rice algorithm. It uses Golomb-Rice codes for entropy coding. The sequence of prediction errors is divided into blocks. Each block is compressed using a two-pass algorithm. In the first pass the best coding method for the whole block is determined. In the second pass, output of the marker of the selected coding method as a side information is done along with prediction errors encoded.

The coding methods include:
	* Golomb-Rice codes of a chosen rank
	* Unary code for transformed pairs of prediction errors
	* Fixed-length natural binary code if the block is found to be incompressible
	* Signaling to the decoder empty block if all prediction errors are zeroes
3. Consultative Committee for Space Data Systems: Lossless Data Compression.
CCSDS Recommendation for Space Data System Standards,
CCSDS 121.0-B-2, Blue Book, May 2012.

|  |
| --- |
|  **Data Template 7.42 - Grid point and spectral data - CCSDS recommended lossless compression.** |
| Octet No. | Contents |
| 6 - nn | CCSDS recommended standard for lossless data compression code stream |

**New entry in Code table 5.0 – Data representation template number**

|  |  |
| --- | --- |
| Octet No. | Contents |
| 42 | Grid point and spectral data - CCSDS recommended lossless compression. |