Dataset Interoperability Recommendations for Earth Science: Part 3

Status of this Memo

This memo provides information to the NASA Earth Science Data System (ESDS) community. This memo does not specify an ESDS standard of any kind. Distribution of this memo is unlimited.

Change Explanation

This document is not a revision to an earlier version.

Copyright Notice

Suggested Citation

Leonard, P.J.T., A. Jelenak, C.S. Zender, E.M. Armstrong, W.E. Baskin, M.J. Brodzik, J.E. Johnson, S.J.S. Khalsa, J.E., W.-H. Li, Y. Wei, M. Brewer, A. Doyle, A. FitzGerrell, C. Lynnes, D.F. Moroni, E. O'Sullivan, B.V. Peters, H.K. Ramapriyan, and V.M. Tsontos. (2024). Dataset Interoperability Recommendations for Earth Science: Part 3. NASA Earth Science Data and Information Systems Standards Coordination Office. https://doi.org/10.5067/DOC/ESCO/ESDS-RFC-054v1.

Abstract

This document contains twelve new recommendations officially adopted by the NASA Earth Science Data System (ESDS) Dataset Interoperability Working Group (DIWG). They are the continuation of the DIWG recommendations published as ESDS-RFC-028 [1] and ESDS-RFC-036 [2] with the same goal of improving the interoperability of Earth Science data product files. The DIWG recommendations here represent best practice guidance covering a diverse range of issues, including flag variables, fill values, projection ellipsoids, geodetic datums, variables in GeoTIFF files, coordinate pairs, valid data ranges, the referencing of documentation, time in seconds since, and the internal compression of variables. Some of the recommendations capture already prevailing community practices while others clarify or simplify among several possible options.

Table of Contents

1 Introduction

The Earth Science Data System Working Groups (ESDSWG) is a NASA organization established under the auspices of NASA Headquarters in 2004. The chartered role of the ESDSWG focuses on the exploration and development of recommendations derived from pertinent community insights of NASA's heterogeneous and distributed Earth Science data systems.

The purpose of the ESDS Dataset Interoperability Working Group (DIWG) is to produce recommendations that make Earth Science data products work better with the tools and services that are commonly used by the community, and to make these products more intercomparable, combinable, and extendable.

This document presents 12 new recommendations from the DIWG regarding dataset interoperability.

The entire set of DIWG Recommendations, both those formally published by ESCO and those DIWG-approved but not yet formally published by ESCO, can be found at [3].

2 Dataset Interoperability Recommendations for Earth Science: Part 1

The DIWG Recommendations Part 1 document (ESDS-RFC-028, [1]) was first released in 2016 (Version 1.0) and was most recently revised in 2018 (Version 1.3). This document includes 12 DIWG Recommendations numbered 2.1 through 2.12[1]:

2.1 Maximize HDF5/NetCDF-4 Interoperability via API Accessibility
2.2 Include Basic CF Attributes
2.3 Use CF "bounds" Attribute
2.4 Verify CF Compliance
2.5 Distinguish Clearly Between HDF and NetCDF Packing Conventions
2.6 When to Employ Packing Attributes
2.7 Mapping Between ACDD and ISO
2.8 Make HDF5 Files NetCDF-4-Compatible and CF-Compliant Within Groups
2.9 Include Time Dimension in Grid Structured Data
2.10 Order Dimensions to Facilitate Readability of Grid Structure Datasets
2.11 Consider "Balanced" Chunking for 3-D Datasets in Grid Structures
2.12 Include Datum Attributes for Data in Grid Structures

[1] The unusual numbering scheme for the DIWG Recommendations resulted from other documents citing each DIWG Recommendation by its section number in the relevant DIWG Recommendations document (e.g., the very first DIWG Recommendation in the very first DIWG Recommendations document is cited as DIWG Recommendation 2.1), and this practice eventually became so commonplace that the DIWG has also adopted it.

3 Dataset Interoperability Recommendations for Earth Science: Part 2

The DIWG Recommendations Part 2 document (ESDS-RFC-036, [2]) was first released in 2019 (Version 1.0) and was most recently revised in 2020 (Version 1.2). This document includes 11 DIWG Recommendations numbered 3.1 through 3.11[1]:

3.1 Character Set for User-Defined Group, Dataset, and Attribute Names
3.2 Consistent units Attribute Value for Variables Across One Data Collection
3.3 Use the units Attribute Only for Variables with Physical Units
3.4 Include Time Coordinate in Swath Structured Data
3.5 Keep Coordinate Values in Coordinate Variables
3.6 Include Georeference Information with Geospatial Coordinates
3.7 Not-a-Number (NaN) Value
3.8 Standardize File Extensions for HDF5/NetCDF Files
3.9 Ensure Granule's Filename Uniqueness Across Different Dataset Releases
3.10 Adopt Semantically Rich Dataset Release Identifiers
3.11 Date-Time Information in Granule Filenames

4 New Dataset Interoperability Recommendations for Earth Science

Presented here are 12 new DIWG Recommendations numbered 4.1 through 4.12[1].

4.1 Attach the CF flag_values and/or flag_masks Attributes Along With the CF flag_meanings Attribute to Each Flag Variable

Recommendation:

Attach the CF flag_values and/or flag_masks attributes along with the CF flag_meanings attribute to each flag variable in an Earth Science data product following the recommendations in the CF Metadata Conventions documentation [4]. The choice of which to use depends upon the use case.

Recommendation Details:

A common practice in Earth Science data products is to associate a flag variable with a science variable to describe the condition(s) associated with each value of the science variable, and making use of the CF flag attributes in such cases can be very helpful. We recommend that the CF flag_values and flag_meanings attributes be attached to the flag variable in cases where the conditions indicated by the flag values are mutually exclusive. In cases where a range of conditions (two or more) for each value of the science variable is possible, flag_masks should be used instead of flag_values. In practice, there can be two flag variables for one science variable: one that has flag_values attached to describe the overall condition of each value of the science variable, and another with flag_masks attached that provides additional details regarding several specific conditions.

In rare (and complex) flagging cases, both flag_values and flag_masks can be attached to the same flag variable, and an example of such usage is provided in Section 3.5 of the CF documentation [4].

flag_values/flag_meanings example:

Use flag_values where a single status condition is sufficient to flag each value of the science variable.

For example, a science data variable named

total_column_ozone

with fill value

_FillValue = -999.9

could be accompanied by an associated flag variable (type byte) named

total_column_ozone_flags

to describe the quality of each successful retrieval as an enumerated list of status flags and to explain each occurrence of the fill value. The total_column_ozone_flags variable should have the CF flag_values and flag_meanings attributes attached to specify the flag values and meanings. For example,

flag_values = 0b, 1b, 2b, 3b, 4b, 5b;

and

flag_meanings = "good_sample glint_contamination high_sza non_convergence row_anomaly_error missing_input_data";

Note that there is a simple one-to-one mapping between the values of these two attributes.

Also note that the values of flag_values are scalars of the same data type as the flag variable and must be comma-separated, while the values of flag_meanings are strings and must be space-separated, with an underscore being used as the word separator within each value of flag_meanings, and with the entire list being surrounded by quotes.

flag_masks/flag_meanings example:

Use flag_masks where a number of independent Boolean (binary) conditions using bit field notation are appropriate to describe a possible range of conditions for each retrieval.

For example, a science data variable named

sea_surface_temperature

could be accompanied by an associated flag variable (type byte) named

sea_surface_temperature_flags

to describe the conditions of each retrieval as a set of Boolean status flags. The sea_surface_temperature_flags variable should have the CF flag_masks and flag_meanings attributes attached to specify the flag values and their meanings. For example,

flag_masks = 1b, 2b, 4b, 8b, 16b;

and

flag_meanings = "ocean land ice lake river";

As an example of implementation, a retrieval could have both ocean and detected ice in it, a condition that would be flagged with the value 1b + 4b = 5b (i.e., $2^0 + 2^2$).

To reiterate two key points, the values of flag_masks must be comma-separated and of the same data type as the flag variable, while the value of flag_meanings must be a list, enclosed in quotes, of space-separated strings.

4.2 Avoid Use of the missing_value Attribute

Recommendation:

Avoid use of the missing_value attribute in new Earth Science data products.

Recommendation Details:

The debate regarding whether the missing_value attribute should be deprecated may never officially come to an end, nevertheless, we recommend that this attribute not be used in new Earth Science data products.

Historically, the missing_value attribute has been used both as a scalar (single-value) and as an array (multi-value) in Earth Science data products.

Use the CF _FillValue attribute instead of using missing_value as a scalar attribute.

Use a flag variable with the CF flag_meanings plus flag_values and/or flag_masks attributes attached (as explained in Recommendation 4.1, above) instead of using missing_value as an array attribute.

It is acceptable to continue to use the missing_value attribute in new versions of data products for continuing Earth Science missions and projects, especially in cases where the downstream software specifically makes use of the missing_value attribute. However, it is strongly recommended to include the CF _FillValue attribute, and, where appropriate, flag variables with the CF flag_* attributes attached (i.e., flag_meanings plus flag_values and/or flag_masks).

4.3 Define the Projection Ellipsoid to Match the Reference Datum

Recommendation:

Define the projection ellipsoid to match the reference datum in an Earth Science data product to minimize potential errors in geolocation and reprojection.

Recommendation Details:

When producing geolocated image data derived from satellite-based or airborne remote sensing instruments, we recommend defining the projection ellipsoid to be the same as the datum used by the remote sensing system to define geodetic latitude and longitude. Specific details depend on the selected file format and metadata conventions. Examples provided in this recommendation use GeoTIFF terminology, but the recommendation also applies to other formats and metadata conventions. For example, when using GeoTIFF to represent content in a Projected Coordinate Reference System (PCRS), the projection ellipsoid should be the same as the Geodetic Reference Frame (datum) used by the remote sensing system. For many currently operating satellite instruments, the reported geolocation is referenced to the World Geodetic System (WGS) 1984 datum. Airborne instruments that are geolocated using GPS instruments are also referenced to WGS 84. When geolocated data from one of these instruments are used to create derived geophysical products, data producers may choose a PCRS that includes a map projection based on a reference ellipsoid. To ensure maximum interoperability when transforming such data products, we recommend choosing the PCRS map projection ellipsoid to match the underlying Geodetic Reference Frame (datum). This will minimize potential for geolocation error with overlays of related geolocated information such as coastlines or comparison data products. We use, in the following discussion, the ISO 19111 terminology described in Section 2.1.2 and Appendix A of [5].

Image transformation of projected data products may require a coordinate conversion or a coordinate transformation. A coordinate conversion is a change of coordinates from one Coordinate Reference System (CRS) to another. The CRSs can be based on the same datum, or, if the datums are different, no algorithm is applied to transform the coordinates of one datum to the other. Since coordinate conversions are considered to be exact, there is no loss of positional accuracy when a coordinate conversion is performed without transforming differing datum coordinates [5]. A coordinate transformation is a change of coordinates from one CRS to another, in which the CRSs are based on different datums. In this case, a coordinate transformation algorithm is applied to convert the coordinates of one CRS to conform to the datum of the other CRS. (Further discussion and a case study example are included in Section 2 of [6].)

It is possible to properly encode both the PCRS map projection ellipsoid and the Geodetic Reference Frame datum in, for example, GeoTIFF metadata. However, some software packages may either incorrectly assume they are the same, or require that they be the same, in order to perform accurate coordinate conversions. Depending on spatial resolution of the image content, the effects of performing an incorrect conversion and/or transformation may not be visually apparent, or may only be apparent if the data include an obvious feature like a subtly shifted coastline when overlaid with independently-derived coastline vectors. A detailed example depicting NASA Operation Ice Bridge flightlines on incorrectly transformed NASA Blue Marble imagery can be found in [6].

Defining the PCRS map projection ellipsoid to match the underlying Geodetic Reference Frame ensures that software packages making this assumption will do the right thing, and will eliminate the time that users might otherwise have to spend to direct the software to only perform the requisite coordinate conversion or coordinate transformation.

In the case of GeoTIFF, this recommendation may become obsolete, given the more specific user-defined details articulated in the OGC GeoTIFF Standard v1.1 [7]. We note that section B.2.3 of [7] explicitly acknowledges historical examples that have used a spherical projection ellipsoid but discourages the use of spheroids for modern applications. Given the relative recentness of this standard, it remains to be seen in practice how closely software packages adhere to encoded projection ellipsoids and reference datums when performing coordinate conversions and transformations.

4.4 Include Only One Science Variable per GeoTIFF File

Recommendation:

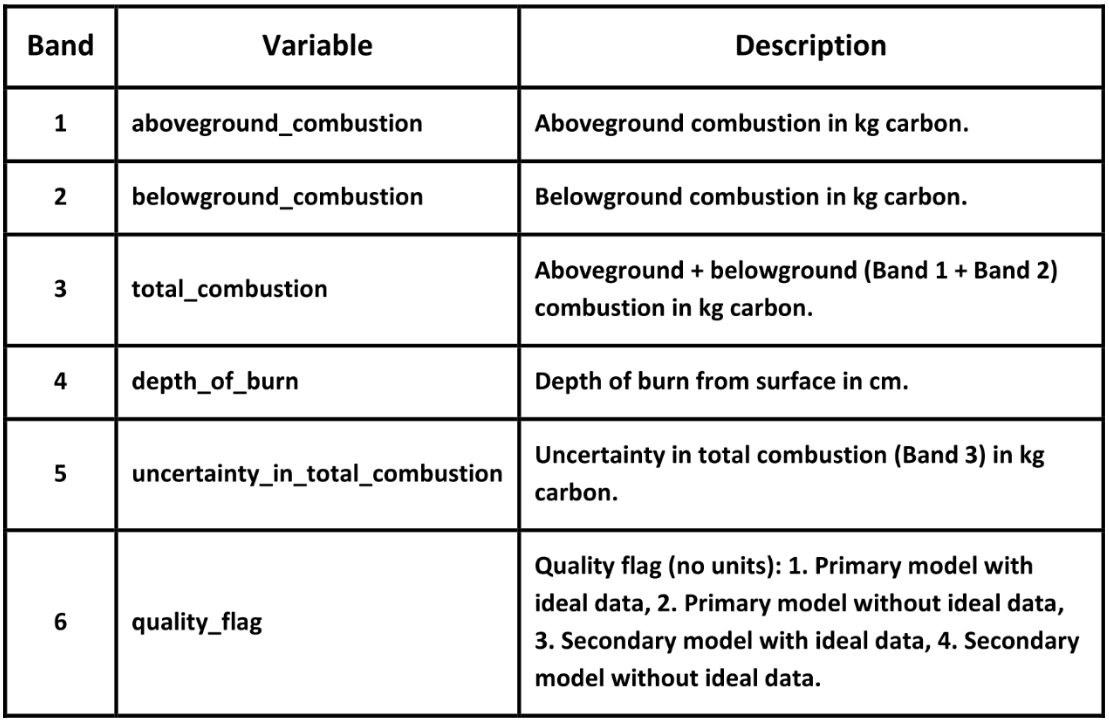
Include only one science variable per GeoTIFF file.

Recommendation Details:

The GeoTIFF format was initially developed during the early 1990s with the objective being to leverage a mature, platform-independent, lossless file format (TIFF) by adding the metadata required for describing and using geographic image data [7]. As defined in Section 4.1 of [7], the term "band" is used in the GeoTIFF format to represent a "range of wavelengths of electromagnetic radiation that produce a single response by a sensing device". A multi-band GeoTIFF file is intended to represent a georeferenced multispectral image, with each band representing a different spectral range, allowing for advanced analysis in remote sensing and GIS applications.

With the ease of use and strong support from both commercial and open-source software [8], the GeoTIFF format has been widely adopted by the Earth Science community to store data beyond the original intended usage of the format. With tools (e.g., Rasterio) that make exporting multi-dimensional arrays into GeoTIFF files easy, some data producers tend to pack multiple different variables as different bands in a GeoTIFF file, even though these different variables do not necessarily represent the response from measurements made in a specific wavelength range.

One such example is the “ABoVE: Burned Area, Depth, and Combustion for Alaska and Canada, 2001-2019” dataset [9] archived at the ORNL DAAC. Table 1 provides a summary of the six "bands" stored in the “ABoVE: Burned Area, Depth, and Combustion for Alaska and Canada, 2001-2019” **combustion_depth** GeoTIFF files. None of these "bands" represents a single response of a sensing device made over a range of wavelengths, which is a major deviation from the intended usage of the GeoTIFF format, and, thus, impairs interoperability.



Band	Variable	Description
1	aboveground_combustion	Aboveground combustion in kg carbon.
2	belowground_combustion	Belowground combustion in kg carbon.
3	total_combustion	Aboveground + belowground (Band 1 + Band 2) combustion in kg carbon.
4	depth_of_burn	Depth of burn from surface in cm.
5	uncertainty_in_total_combustion	Uncertainty in total combustion (Band 3) in kg carbon.
6	quality_flag	Quality flag (no units): 1. Primary model with ideal data, 2. Primary model without ideal data, 3. Secondary model with ideal data, 4. Secondary model without ideal data.

Table 1. A summary of the six "bands" stored in the “ABoVE: Burned Area, Depth, and Combustion for Alaska and Canada, 2001-2019” **combustion_depth** GeoTIFF files.

We Discourage Packing Multiple Different Variables in a Multi-Band GeoTIFF File:

Although it is fairly easy for data producers to create multi-band GeoTIFF files like in the example presented in Table 1, it may be challenging for general users to fully understand and properly use such files. This is mainly due to the lack of an agreed-upon approach to tag GeoTIFF files with metadata (e.g., science variable name and units) for individual bands. Even though there are libraries (like GDAL) that support adding custom-defined metadata tags to individual bands of GeoTIFF files, many tools and applications still lack the capability to add or use these custom-defined metadata tags. Such GeoTIFF files become less self-descriptive, thus decreasing their interoperability. Also, the different data variables must all be of the same data type (e.g., Int16 or Float32, not a mixture of Int16 and Float32) once being physically stacked as multiple bands in a GeoTIFF file.

Some Advice Regarding Storing Complex Data:

Rather than using GeoTIFF to store data that have multiple variables and/or dimensions, please consider using formats like HDF5 and/or netCDF-4, which have established community standards/conventions to embed variable- and file-level metadata to make such files self-descriptive and interoperable. The data from such files can subsequently be used to generate single-variable GeoTIFF files.

If GeoTIFF is the preferred format, then please include only one science variable per GeoTIFF file, and consider an approach, such as GDAL's Virtual Dataset (VRT) format, to "virtually" aggregate multiple single-variable GeoTIFF files together into a metadata-rich data file/asset without having to change their native data types.

4.5 Indicate in CRS Metadata the Order of Elements in Horizontal Coordinate Pairs

Recommendation:

Indicate in CRS metadata the order of latitude and longitude in coordinate pairs in an Earth Science data product.

Recommendation Details:

There is no universal agreement regarding the order of horizontal coordinate pairs (i.e., (longitude, latitude) vs. (latitude, longitude)) in Earth Science data products. Axis ordering may be specified in the full description of the Coordinate Reference System (CRS) as given in a registry such as EPSG. If the order is not specified in a registered CRS, or the CRS is not in a registry, we recommend using the optional axis order keyword in the well-known text (WKT) representation of a CRS (ISO 19162:2019 [10]). The order keyword can be added after the mandatory direction keyword as shown in this example:

```
AXIS["longitude",east,ORDER[1]],

AXIS["latitude",north,ORDER[2]],
```

The DIWG recommends that CRS be included in all Earth Science dataset granules that contain geospatial coordinates, and that WKT should be included whenever possible (see Recommendation 3.6 of [2], and Section 5.4 below).

But if a CRS is not specified using WKT, then we recommend that ISO 6709 [11] be followed, which states that the following shall apply when no CRS is provided:

1. Within a coordinate tuple, the latitude value shall precede the longitude value.
2. Latitudes on or north of the equator shall be positive, latitudes south of the equator shall be negative.
3. Longitudes on or east of the prime meridian shall be positive, longitudes west of the prime meridian shall be negative. The 180th meridian shall be negative[2]. The prime meridian shall be Greenwich.

[2] It should be noted that the range in longitude must be -180° to +180° if ISO 6709 is followed, but, apart from this one case, the DIWG is not mandating any particular valid range for longitude.

4.6 Make a Variable's Valid Data Range Consistent Within Each Product Release

Recommendation:

The valid data range for each variable in an Earth Science data product should be made consistent within each product release, and should not vary file-to-file within a given product release.

Recommendation Details:

There are cases of published Earth Science data products with valid data ranges for some variables (specified via the CF valid_min and valid_max attributes, or via the CF valid_range attribute)[3] that vary file-to-file, based on the actual data range for each particular variable within each particular product file, which is an approach that we do not agree with.

The valid data range for each variable in an Earth Science data product should be made consistent within each product release, and should not vary file-to-file within a given product release.

The valid data range for any given variable should be based upon the relevant physics for that variable, and, possibly, constraints placed upon the variable based on the characteristics of the sensor(s) that collected the data. The relevant physics should not vary file-to-file. The characteristics of the sensor(s) can change with time, in which case we recommend that the widest valid data range consistent with the changing characteristics of the sensors(s) be used for each variable within a given product release.

We note that the choice of the valid data range for any given variable places a constraint on the choice of the fill value for that variable, because the fill value must be a number outside of the valid data range (as explained in Recommendation 4.8, below).

[3] This recommendation is intended to be a general recommendation that applies to all Earth Science data products, even though the example implementation presented here focuses on the CF Metadata Conventions.

4.7 Make a Variable's Valid Data Range Useful

Recommendation:

The valid range for each variable in an Earth Science data product should put useful constraints on the data.

Recommendation Details:

Declaring the valid range of a variable's data according to the CF metadata conventions is part of an earlier DIWG recommendation (see Recommendation 2.1 of [1]). The data value range can be specified either by two CF attributes, valid_min and valid_max, or via the valid_range CF attribute[3]. Only one of these approaches should be used for a given variable.

The data ranges declared using these attributes are dependent on the type of data and their intended application, and should be chosen to place meaningful constraints on the possible data values. The CF metadata conventions require that any data value representing missing data or the variable's fill value must be excluded from the valid data range.

Generic range values are discouraged unless the actual data range is poorly understood. For example, we strongly discourage using the limits of a specific computer data type (e.g., floating-point single or double precision) as the valid range. If the valid range is poorly understood for a particular variable, then it would be best not to include the valid range attribute(s) for that variable.

A useful valid range allows scientists and other users to filter out values that violate physics or known characteristics of the sensor. It also allows visualization programs to either ignore such points or display them with a special style to warn users of the constraint violation.

4.8 Use a Number Outside of the Valid Data Range for a Variable's Fill Value

Recommendation:

The fill value of a (non-string) variable should be a number outside its valid data range in an Earth Science data product.

Recommendation Details:

The CF _FillValue attribute is used to indicate missing or invalid data for a variable[3]. Also, the value of the CF _FillValue attribute should match the actual fill value used for the variable in the file.

For non-string variables, the value of the CF _FillValue attribute should be a mathematically valid number that lies outside the valid range for a variable. Please note that NaN (Not-a-Number) is neither a number nor is it mathematically valid, and, thus, should not be used as the fill value (see Recommendation 3.7 of [2]).

For string variables, see Recommendation 4.12 below.

Using zero as the fill value should be avoided, because zero looks too much like a physically realistic value, and this can be confusing to the product users.

There should only be one fill value per variable. We recommend using a quality flag variable with the CF flag_meanings plus flag_values and/or flag_masks attributes attached to explain the various reasons for using the fill value (see Recommendation 4.1, above), instead of using several special values in the variable.

4.9 Use DOIs for Referencing Documentation

Recommendation:

A space-separated list of documentation DOIs should be used in the CF references attribute in Earth Science data products, both globally and for specific variables.

Recommendation Details:

The CF references attribute is useful for storing information regarding documentation in an Earth Science data product. The CF references attribute can exist both globally and at the variable level. The most concise way to reference a document is via its DOI. We suggest that a space-separated list of documentation DOIs should be used in the CF references attribute in Earth Science data products. Use of the URL form of the DOI is strongly recommended. Also, URLs of relevant documents that do not have DOIs can be used in the CF references attribute, though it should be noted that non-DOI URLs do not have the same level of permanency as DOI URLs.

4.10 Use Double Precision When Archiving Time in Seconds Since a Specific Epoch

Recommendation:

Use double precision when archiving time in seconds since a specific epoch in an Earth Science data product.

Recommendation Details:

Earth Science data products must preserve time-related information with sufficient precision to resolve all timescales relevant to the data itself, to other data with which it may be intercompared, and to conventions for the numeric representation of time, such as Coordinated Universal Time (UTC). Geoscientific datasets commonly report time in intervals (such as seconds) measured from a particular epoch. Resolving one second on the 50-year-plus timescale from the UNIX/POSIX epoch (00:00:00 UTC on 1 January 1970) to the present day can require up to ten significant digits of temporal resolution, whereas the IEEE-754 single-precision (32 bit) floating point representations preserves at most seven significant digits. Resolving time to the nearest microsecond can require up to six more digits, for a total of sixteen digits, approximately the maximum precision of an IEEE-754 double-precision (64-bit) floating point number. Therefore, preserving sufficient temporal precision to label, store, and intercompare geoscientific data requires double-precision storage.

The most straightforward way of implementing this recommendation in an Earth Science data product is to make use of a double-precision time variable.

A somewhat less straightforward, but perfectly legitimate, way of implementing this recommendation in an Earth Science data product is to

1) include in each Earth Science data product file a double-precision **time of reference**[4] in seconds since a specific epoch

and

2) provide within the Earth Science data product file the time of each individual observation via a 4-byte (perhaps scaled) integer w.r.t. the double-precision **time of reference**[4].

Combining 1) and 2) results in time in seconds since a specific epoch in double precision. The limitation of this approach depends on how many digits to the right of the decimal place must be included. If time is to be reported to the nearest 0.01 second, then a 4-byte integer would require a scale factor of 0.01 for conversion to seconds, in which case a 4-byte integer could only hold one day's worth of information w.r.t. **the time of reference**[4], and would limit this approach to sub-daily and daily product files.

[4] The most commonly used **time of reference** is the time at 00:00:00 UTC on the date of measurement.

4.11 Use Only Officially Supported Compression Filters on NetCDF-4-Compatible Data

Recommendation:

Only compression filters that are officially supported by a default installation of the current netCDF-4 software distribution should be used in Earth Science data products in netCDF-4-compatible formats.

Recommendation Details:

Only compression filters that are officially supported by a default installation of the netCDF-4 software should be used in interoperable Earth Science data products in netCDF-4-compatible formats. The allowed compression filters currently[5] include DEFLATE, bzip2, zstandard, and blosc (a description of netCDF-4 filter support can be found at [12]). While netCDF-4 has enabled access to all HDF5 compression filters starting from version 4.7.0, use of any filter not installed by netCDF-4 is strongly discouraged. This is because the identification and invocation methods for these HDF5 filters are obscure (involving five digit IDs) and non-portable (no guarantees client software will be able to decompress them). Use of the shuffle filter is not prohibited since it is not a compression filter and is supported by the netCDF-4 default installation. Combining shuffle with compression filters can noticeably improve the data compression ratio in many cases.

[5] At around the time of publication of this document.

4.12 Use the ASCII Null Character (0x00) as the Fill Value for String Data and Metadata

Recommendation:

Use the ASCII null character (0x00) as the fill value for string data and metadata in Earth Science data products.

Recommendation Details:

Inaccurate values of string data or metadata, whether static or dynamic, should be avoided in an Earth Science data product, and the simplest solution is to replace such values with the appropriate fill value.

We recommend that the null character be used as the fill value for string data and metadata in Earth Science data products.

The null character can be accessed in a variety of ways:

- The ASCII null character is **0x00**.
- The null character is Unicode code point **U+0000**.
- The Fortran 90 programming language command **achar(0)** returns the null character.
- The netCDF-4 library Fortran 90 interface variable **nf90_fill_char** is the null character.
- The C programming language represents the null character as **\0**.
- The netCDF-4 library C interface pre-preprocessor tokens, **NC_FILL_CHAR** and **NC_FILL_STRING**, both represent the null character. Use the former when the string value is a fixed-length array of type NC_CHAR. Use the latter when the string value is stored in a dynamic array of type NC_STRING.

5 Families of Dataset Interoperability Recommendations for Earth Science

With the advent of the new set of DIWG Recommendations presented in Section 4, it became apparent that there are three Families and one Super Family of Dataset Interoperability Recommendations for Earth Science data products.

5.1 "Valid Range" Family of Recommendations

There are three DIWG Recommendations related to the CF valid_* attributes:

1. (The valid_* part of) Recommendation 2.2 "Include Basic CF Attributes."
2. Recommendation 4.6, "Make a Variable's Valid Data Range Consistent Within Each Product Release."
3. Recommendation 4.7, "Make a Variable's Valid Data Range Useful."

These three recommendations should be considered together when developing an Earth Science data product.

5.2 "Fill Value" Family of Recommendations

There are two DIWG Recommendations directly related to the _FillValue attribute:

1. (The _FillValue part of) Recommendation 2.2, "Include Basic CF Attributes."
2. Recommendation 4.8, "Use a Number Outside of the Valid Data Range for a Variable's Fill Value."

The DIWG has agreed that these two Recommendations alone are enough to represent a Family of Recommendations.

However, there are three additional Recommendations that are related to fill values:

1. Recommendation 3.7, "Not-a-Number (NaN) Value," which suggests that NaN should not be used as the fill value.
2. Recommendation 4.2, "Avoid Use of the missing_value Attribute," which suggests that the _FillValue attribute should be used instead of the missing_value attribute.
3. Recommendation 4.12, “Use the ASCII Null Character (0x00) as the Fill Value for String Data and Metadata," which applies to strings.

These five recommendations should be considered together when developing an Earth Science data product.

5.3 "Valid Range and Fill Value" Super Family of Recommendations

The CF valid_* attributes are also related to the _FillValue attribute, because the fill value must be a number outside of the valid data range, and so the "Valid Range" Family of Recommendations and the "Fill Value" Family of Recommendations together represent a Super Family of DIWG Recommendations.

5.4 Georeferencing Family of Recommendations

There are four DIWG Recommendations related to georeferencing:

1. Recommendation 2.12, "Include Datum Attributes for Data in Grid Structures."
2. Recommendation 3.6, "Include Georeference Information with Geospatial Coordinates."
3. Recommendation 4.3, "Define the Projection Ellipsoid to Match the Reference Datum."
4. Recommendation 4.5, "Indicate in CRS Metadata the Order of Elements in Horizontal Coordinate Pairs."

These four recommendations should be considered together when developing an Earth Science data product.

6 Endorsement of Data Product Development Guide for Data Producers

The Data Product Development Guide for Data Producers (DPDG) [13] is an aid aimed primarily at data producers who develop Earth Science data products that are to be archived at an EOSDIS DAAC, though product developers who do not archive their products at an EOSDIS DAAC may also find the DPDG to be useful.

The DIWG endorses the entire DPDG, especially Appendix D on global attributes and Appendix E on variable-level attributes.

Also, the section of the DPDG on cloud-optimized formats and services, which is new in Version 2.0 of the DPDG [13], should be helpful for the development of data products for use in a cloud environment.

7 References

[1] Charles S. Zender, Peter J.T. Leonard, et al., "Dataset Interoperability Recommendations for Earth Science," 2016, https://www.earthdata.nasa.gov/s3fs-public/imported/ESDS-RFC-028v1.3.pdf.

[2] Aleksandar Jelenak, Peter J.T. Leonard, et al., "Dataset Interoperability Recommendations for Earth Science: Part 2," 2020, https://www.earthdata.nasa.gov/s3fs-public/imported/ESDS-RFC-036v1.2.pdf.

[3] DIWG, "Dataset Interoperability Recommendations for Earth Science," https://wiki.earthdata.nasa.gov/display/ESDSWG/Dataset+Interoperability+Recommendations+for+Earth+Science.

[4] Climate and Forecast Metadata Conventions, Section 3.5 Flags, http://cfconventions.org/Data/cf-conventions/cf-conventions-1.11/cf-conventions.html#flags.

[5] E. Iliffe, J. and Lott, R., "Datums and Map Projections for Remote Sensing, GIS and Surveying, 2nd Edition," 2008 May 14, Whittles Publishing, Dunbeath, Scotland.

[6] Brodzik, M. J., B. Billingsley, T. Haran, B. Raup, and M. H. Savoie, "EASE-Grid 2.0: Incremental but Significant Improvements for Earth-Gridded Data Sets," 2012, ISPRS International Journal of Geo-Information, Volume 1, Number 1, Pages 32-45, https://doi.org/10.3390/ijgi1010032.

[7] Open Geospatial Consortium, "OGC GeoTIFF Standard, Version 1.1," 2019 September 14, OGC Document Number 19-008r4, http://docs.opengeospatial.org/is/19-008r4/19-008r4.html.

[8] ESDIS Standards Coordination Office, "ESCO Recommended Standard, OGC GeoTIFF Standard, Version 1.1," 2019, https://www.earthdata.nasa.gov/esdis/esco/standards-and-practices/geotiff and https://www.earthdata.nasa.gov/s3fs-public/imported/ESDS-RFC-040v1.1.pdf.

[9] Potter, S., S. Veraverbeke, X.J. Walker, M.C. Mack, S.J. Goetz, J.L. Baltzer, C. Dieleman, N.H.F. French, E.S. Kane, M.R. Turetsky, E.B. Wiggins, and B.M. Rogers. 2022. ABoVE: Burned Area, Depth, and Combustion for Alaska and Canada, 2001-2019. ORNL DAAC, Oak Ridge, Tennessee, USA, https://doi.org/10.3334/ORNLDAAC/2063.

[10] International Organization for Standardization, ISO 19162:2019, Geographic information, Well-known text representation of coordinate reference systems, 2019-07.

[11] International Organization for Standardization, ISO 6709:2022, Standard representation of geographic point location by coordinates, 2022-09.

[12] Unidata Network Common Data Form (NetCDF), "NetCDF-4 Filter Support," https://docs.unidata.ucar.edu/netcdf-c/current/filters.html.

[13] Ramapriyan, H. K., P. J. T. Leonard, E. M. Armstrong, et al., "Data Product Development Guide (DPDG) for Data Producers Version 2.0," 2024 July, NASA Earth Science Data and Information System Standards Coordination Office, https://doi.org/10.5067/DOC/ESCO/RFC-041VERSION2 and https://www.earthdata.nasa.gov/esdis/esco/standards-and-practices/data-product-development-guide-for-data-producers.

8 Authors and Contact Information

8.1 DIWG Leadership

Peter J.T. Leonard
peter.j.leonard@nasa.gov
https://orcid.org/0009-0002-8007-5784

Aleksandar Jelenak
ajelenak@hdfgroup.org
https://orcid.org/0009-0001-2102-0559

Charles S. Zender
zender@uci.edu
https://orcid.org/0000-0003-0129-8024

8.2 DIWG Recommendation Suggestors, Drafters, and Re-Drafters

(Excludes Anyone Listed Previously)

Edward M. Armstrong
edward.m.armstrong@jpl.nasa.gov
https://orcid.org/0000-0002-5595-9353

Walter E. Baskin
walter.e.baskin@nasa.gov
https://orcid.org/0000-0002-2241-3266

Mary J. Brodzik
brodzik@colorado.edu
https://orcid.org/0000-0002-2544-659X

James E. Johnson
james.e.johnson-1@nasa.gov
https://orcid.org/0009-0005-2436-0216

Siri Jodha S. Khalsa
khalsa@colorado.edu
https://orcid.org/0000-0001-9217-5550

Wen-Hao Li
whl52059@gmail.com
https://orcid.org/0000-0001-6680-865X

Yaxing Wei
weiy@ornl.gov
https://orcid.org/0000-0001-6924-0078

8.3 DIWG Recommendation Major Commentators

(Excludes Anyone Listed Previously)

Marty Brewer
brewer@remss.com
https://orcid.org/0009-0002-0311-8453

Allan Doyle

Amy FitzGerrell
amy.fitzgerrell@colorado.edu
https://orcid.org/0000-0001-7834-5461

Christopher Lynnes
https://orcid.org/0000-0001-6744-3349

David F. Moroni
david.f.moroni@gmail.com
https://orcid.org/0000-0003-2994-557X

Ewan O'Sullivan
eosullivan@cfa.harvard.edu
https://orcid.org/0000-0002-5671-6900

Byron V. Peters
byron.v.peters@nasa.gov
https://orcid.org/0000-0003-2130-2756

Hampapuram K. Ramapriyan
hampapuram.ramapriyan@ssaihq.com
https://orcid.org/0000-0002-8425-8943

Vardis M. Tsontos
https://orcid.org/0000-0002-1723-0860

Appendix A - Acronyms

Acronym	Description
ABoVE	Arctic-Boreal Vulnerability Experiment
ACDD	Attribute Convention for Data Discovery
API	Application Program Interface
ASCII	American Standard Code for Information Interchange
CF	Climate and Forecast Metadata Conventions
CRS	Coordinate Reference System
DAAC	Distributed Active Archive System
DIWG	Dataset Interoperability Working Group
DOI	Digital Object Identifier
DPDG	Data Product Development Guide for Data Producers
EOSDIS	Earth Observing System Data and Information System
EPSG	European Petroleum Survey Group
ESCO	ESDIS Standards Coordination Office
ESDIS	Earth Science Data and Information System
ESDS	Earth Science Data System
ESDSWG	Earth Science Data System Working Groups
ESO	EDSIS Standards Office (renamed to ESCO)

Acronym	Description
GDAL	Geospatial Data Abstraction Library
GeoTIFF	Georeferenced Tagged Image File Format
GIS	Geographic Information Systems
GPS	Global Positioning System
HDF	Hierarchical Data Format
HDF5	Hierarchical Data Format Version 5
IEEE	Institute of Electrical and Electronics Engineers
ISO	International Organization for Standardization
NaN	Not-a-Number
NASA	National Aeronautics and Space Administration
NetCDF	Network Common Data Form
NetCDF-4	Network Common Data Form Version 4
OGC	Open Geospatial Consortium
ORCID	Open Researcher and Contributor ID
ORNL	Oak Ridge National Laboratory
PCRS	Projected Coordinate Reference System
POSIX	Portable Operating System Interface
RFC	Request For Comments
TIFF	Tagged Image File Format
URL	Universal Resource Locator
UTC	Coordinated Universal Time
VRT	Virtual Raster Format
WGS	World Geodetic System
WGS 84	World Geodetic System 1984
WKT	Well-Known Text

Appendix B - Glossary

Term	Description
blosc	A high-performance, blocking, and shuffling compression library designed for binary data.
bzip2	A free and open-source file compression program that uses the Burrows–Wheeler algorithm.
DEFLATE	A data compression algorithm based on Huffman coding and LZ77 compression.
Geodetic Datum	A geodetic datum is a global reference frame for precisely representing the position of locations on Earth by means of geodetic coordinates. The plural form of the term geodetic datum is geodetic datums.
LZ77	A lossless compression algorithm invented by Abraham Lempel and Jacob Ziv in 1977.
Rasterio	A Python library that facilitates reading, writing, and analyzing geospatial raster data.
shuffle filter	An HDF5 filter that rearranges the bytes in integer data in a way that increases redundancy, thereby improving the effectiveness of the chosen compression algorithm.
Unicode	A universal character encoding standard that assigns a unique number (a code point) to every character in every language, including alphabets, symbols, and ideographs.
UNIX Time	UNIX Time is a date and time representation that originated as the system time in the UNIX operating system.
zstandard	A fast, open-source compression algorithm developed by Yann Collet at Facebook that offers a good balance between compression speed and ratio.