AICCA: the AI-driven Cloud Classification Atlas

Takuya Kurihana

Version: September 30, 2022

Dataset available at https://github.com/RDCEP/clouds#download-aicca-dataset

Contents

1 Overview **1**

2 AICCA: data and outputs **1**
2.1 MODIS Data 2
2.2 AICCA Patch-Level Data 2
2.3 AICCA Grid Cell-Level Data 3
2.4 AICCA Intermediate-Level Data 5

3 Download AICCA **5**
3.1 Globus 5
3.2 Missing dates 6

4 How to read and use AICCA **7**

5 Contact **7**

1 Overview

We release **the AI-driven Cloud Classification Atlas (AICCA)**as a novel cloud classification dataset produced by applying modern self-supervised deep learning methods to identify robust and meaningful clusters of cloud patterns. We apply these self-supervised deep learning methods to 22 years of data from the Moderate Resolution Imaging Spectroradiometer (MODIS) instruments on NASA's Aqua and Terra satellites to produce global-scale AI-generated 42 cloud classes as a standardized science product. AICCA delivers in a compact form (tens of gigabytes of class labels, with high spatial and temporal resolution) information currently accessible only as hundreds of terabytes of multi-spectral images: AICCA enables data-driven diagnosis of patterns of cloud organization, provides insight into their evolution on timescales of hours to decades, and contributes to the democratization of climate research by facilitating access to core data. The preliminary investigation discovers classes based on both cloud morphology and physical properties, and yields unbiased cloud classes free from artificial assumptions that capture the diversity of global cloud types. AICCA is also intended to support studies of the response of clouds to forcing on timescales from hours to decades and to allow data-driven diagnosis of cloud organization and behavior and their evolution over time as CO_2 and temperatures increase.

2 AICCA: data and outputs

The dataset described here, $AICCA_{42}$ (or simply AICCA), provides AI-generated cloud class labels for all ocean clouds sampled by MODIS instruments over their 22 years of operation. The cloud labels (range 1..42) are generated by the rotation-invariant cloud clustering (RICC) method [1] configured to generate 42 clusters, using input imagery subdivided into 128×128 pixel (∼100 km by 100 km)

Table 1: MODIS products used to create the AICCA dataset. As noted in the text, each product name *MOD0X* in the first column refers to both the Aqua (MYD0X) and Terra (MOD0X) products. Source: NASA Earthdata.

Product	Description	Band	Primary Use	Process
MOD02	Shortwave infrared (1.230–1.250 μm)	5	Land/cloud/aerosol properties	Patch
	Shortwave infrared (1.628–1.652 μm)	6	Land/cloud/aerosol properties	
	Shortwave infrared (2.105–2.155 μm)	7	Land/cloud/aerosol properties	
	Longwave thermal infrared (3.660–3.840 μm)	20	Surface/cloud temperature	
	Longwave thermal infrared (7.175–7.475 μm)	28	Cirrus clouds water vapor	
	Longwave thermal infrared (8.400–8.700 μm)	29	Cloud properties	
	Longwave thermal infrared (10.780–11.280 μm)	31	Surface/cloud temperature	
MOD03	Geolocation fields		Latitude and Longitude	QC
MOD06	Cloud mask		Cloud pixel detection	
	Land / Water		Background detection	
	Cloud optical thickness		Thickness of cloud	Evaluation
	Cloud top pressure		Pressure at cloud top	
	Cloud phase infrared		Cloud particle phase	
	Cloud effective radius		Radius of cloud droplet	

patches. Labeled output is provided in two different ways: as the original patches and resampled to $1^\circ \times 1^\circ$ *grid cells*.

2.1 MODIS Data

The MODIS instruments hosted on NASA's Aqua and Terra satellites have been collecting visible to mid-infrared radiance data in 36 spectral bands from 2002 (Aqua) [2] and 2000 (Terra) [3] through 2021. The instruments collect data over an approximately 2330 km by 2030 km *swath* every five minutes, with a spatial resolution of 1 km. AICCA is based on the MODIS Level 1B calibrated radiance product (MOD02). (Note that while NASA uses the prefixes MOD and MYD to distinguish between Terra and Aqua, respectively, for simplicity we use MOD to refer to both throughout this article.) We limit the dataset to the 6 spectral bands most relevant for derivation of physical properties: bands 6, 7, and 20 relate to cloud optical properties, and bands 28, 29, and 31 relate to the separation of high and low clouds and the detection of the cloud phase. For the Aqua instrument, we use band 5 as an alternative to band 6 due to a known stripe noise issue in Aqua band 6 [4]. (See also [1] for more details.) The total number of swath images per band is (12 swath/hour) $\times$ (12 hour/day) $\times$ (365 day/year) $\times$ (20 + 22 years, for Aqua and Terra, respectively) $\approx$ 2.2 million swathes, or a total of 13.2 million swathes across all bands used.

MODIS multispectral data are processed by NASA to yield a variety of products, and we use several others for post-processing or analysis. We use latitude and longitude from the MOD03 geolocation fields to regrid the AICCA patches, and use selected derived physical properties from the MOD06 product to evaluate the cloud classes: four physical parameters related to cloud optical properties and cloud top properties (CTP and CPI). Note that the MOD06 variables are used only as a diagnostic, to evaluate to what extent clustering results associate cloud physical properties in our clusters. They are not included in our RICC training data, so as to free from any assumptions imposed on these data. All data used in generating AICCA are listed in Table 1, and make up an aggregate size of 801 Tb. All MODIS products are made available via the NASA Level-1 and Atmosphere Archive and Distribution System (LAADS), grouped into per-swath files.

2.2 AICCA Patch-Level Data

The AICCA dataset uses all patches from Aqua and Terra MODIS image data during 2000–2021, subject to the constraints that they are (a) disjoint in space and/or time; (b) include no non-ocean pixels, and (c) each include at least 30% cloud pixels. The resulting set is 198 676 800 individual 128×128 pixel ($\sim$100 km by 100 km) ocean-cloud patches (numbered in first element of the AICCA_{42} dataset). For each patch, AICCA_{42} provides the information listed below and in Table 2, a total of 146 Bytes per patch, i.e. a factor of 16 159 $\times$ reduction relative to the raw multispectral imagery.

Table 2: Information provided in AICCA for each 128×128 pixel ocean-cloud patch: metadata that locate the patch in space and time, and indicate whether the patch was used to train RICC; a cloud class label computed by RICC; and a set of diagnostic quantities obtained by aggregating MODIS data over all pixels in the patch.

Variables	Description	Values	Type
Swath	Identifier for source MODIS swath	1	float32
Location	Geolocation index for the upper left corner of patch	2	float32
Timestamp	Time of observation	1	float32
Training	Whether patch used for training	1	binary
Label	Class label assigned by RICC: integer in range $1..k^*$	1	int32
COT_patch	Mean and standard deviation of pixel values in patch	2	float32
CTP_patch	”	”	”
CER_patch	”	”	”
CPI_patch	Number of pixels in patch in {clear-sky, liquid, ice, undefined}	4	int32

- `Source` is either Aqua or Terra;
- `Swath`, `Location`, and `Timestamp` locate the patch in time and space;
- `Training` indicates whether the patch was used for training;
- `Label` is an integer in the range 1..42, generated by the rotation-invariant cloud clustering system configured for 42 clusters, RICC_{42}; and
- `COT_patch`, `CTP_patch`, and `CER_patch`: The mean and standard deviation, across all pixels in the patch, for three MOD06 physical values: cloud optical thickness (COT), cloud top pressure (CTP), and cloud effective radius (CER);
- `CPI_patch`: Cloud phase information (CPI), four numbers representing the number of the 128×128 pixels in the patch that are estimated as clear-sky, liquid, ice, or undefined, respectively.

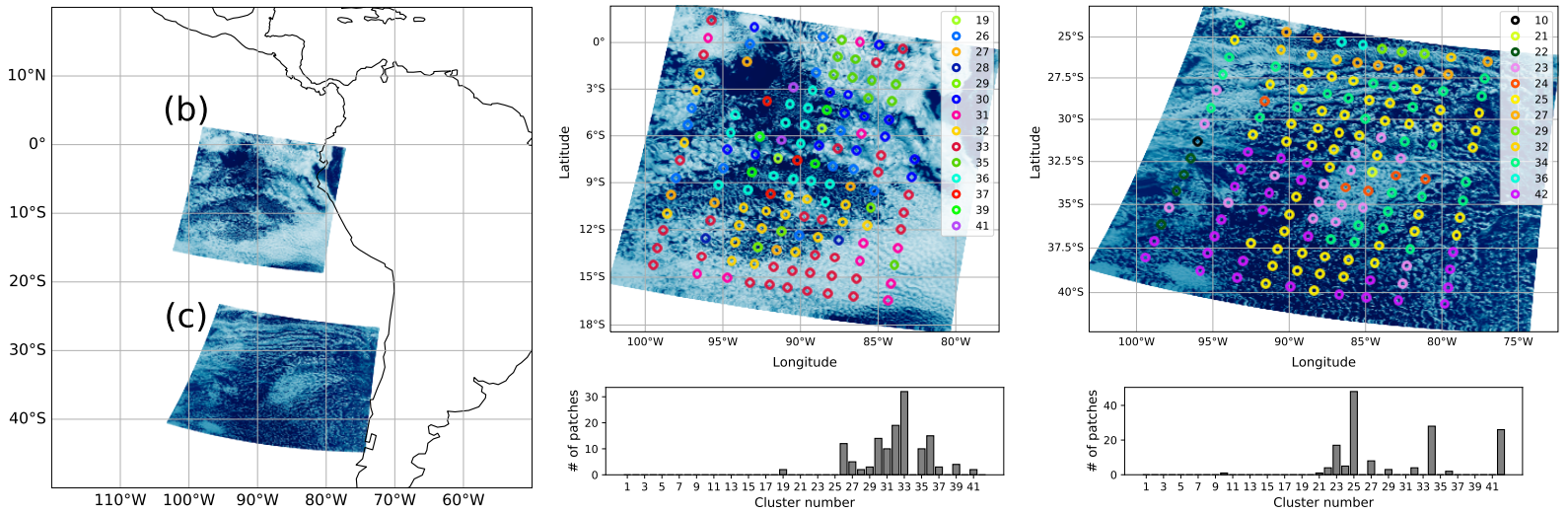
The additional information shown in Table 2 that assists users in understanding each patch is easily extracted from MOD06 by using the patch's geolocation index and timestamp (Location and Timestamp in Table 2) to locate the patch's data in the appropriate MOD06 file. These mean values summarize the average physical characteristic for the patch; the standard deviations provide some indication as to the existence of multiple clouds (especially low- and high-altitude clouds). We do not use the MOD06 multilayered cloud flag.

Output is provided as NetCDF [5] files that combine patches from each MODIS swath into a single file. While AICCA contains no raw satellite data, note that information for each patch in Table 2 includes an identifier for the source MODIS swath and a geolocation index, that allow users to link results with the original MOD02 satellite imagery and other MODIS products. The complete `OC-Patches` set contains around (20 + 22 years) × (365 days/year) × (26 000 patches) ×146 B ≈ 54.2 gigabytes.

The example swathes, as shown in Figure 1a, are both off the coast of Peru: from Terra on January 15, 2003 (Figure 1b) and July 20, 2003 (Figure 1c). Each dot in Figures 1b and 1c denotes a cluster label assigned to a patch with >30% cloud pixels. Visual inspection of both swathes shows both a pronounced *separation* among different cloud textures but also *coherence* when similar textures occur in neighboring pixels.

2.3 AICCA Grid Cell-Level Data

In addition to providing per-patch data, we follow common practice in climate datasets by also providing data organized on a per-latitude/longitude grid cell basis. The second element of the AICCA_{42} dataset spatially aggregates the patch-level class label and diagnostic values at a resolution of $1^\circ \times 1^\circ$, a total of 181 × 360 *grid cells* over the globe. For each resulting data item, AICCA_{42} provides the information listed in Table 3, a total of 32 Bytes:



(a) Geographical location of swathes

(b) Patch labels with 42 clusters

(c) Patch labels with 42 clusters

Figure 1: (a) The location of two swathes off the coast of Peru. (b) 133 ocean-cloud patches from the first swath (18° S to 3° N, 76° S to 104° S) observed by the Terra instrument on January 15, 2003. (c) 147 ocean-cloud patches from the second swath (44° S to 23° S, 72° S to 103° S) observed by the Terra instrument on July 20, 2003. In (b) and (c), each color marker denotes a patch cluster label, in the range 1..42 (not all clusters appear in each swath); the histograms show the distribution of cloud labels. We see that AICCA produces spatially coherent cluster assignments for similar cloud objects: e.g., closed-cell stratocumulus in the lower part of (b) and (c). AICCA also captures richer information than the standard cloud classification, e.g., it separates the different textures of open- and closed-cell stratocumulus clouds into multiple classes.

- `Source` is either Aqua or Terra;
- `Cell` gives a latitude and longitude for the grid cell;
- `Timestamp` locates the grid cell in time;
- `Label_1deg` represents the most frequent class label in the grid cell (an integer in the range 1..42); and
- `COT_1deg`, `CTP_1deg`, `CER_1deg`, and `CPI_1deg` aggregate values for four diagnostic variables, as described in Section 2.3.

The aggregation process uses values from individual days from the Aqua and Terra satellite, a reasonable choice since the MODIS instrument orbits the Earth about once per day. That is, the swaths taken by each satellite instrument generally do not overlap in a daily period. Since a single 2330 km by 2030 km MODIS swath extends across multiple 1 degree by 1 degree grid cells, we extract the latitude and longitude at the center of each `OC-Patch` by using MOD03, and aggregate the information listed in Table 2 to each 1° × 1° grid cell (i.e., the area extending from −0.5° to +0.5° from the grid cell center). To assign a class label to each grid cell on each day, we use the class of the single ocean-cloud patch with the largest overlap with the grid cell. To provide physical properties for each grid cell, we do implement one simplification to reduce the use of computing memory: instead of averaging pixel values within each grid cell, we identify *all* ocean-cloud patches that overlap with the cell, and simply average those patches' mean COT, CTP, and CER values. To assign a cloud particle phase (clear–sky, liquid, ice, or undefined) we use the most frequent phase in the overlapping patches. Grid cells with no clouds are labeled as a missing value.

In some cases, especially at high latitudes, swaths may overlap within a single day. When this occurs, patches with different timestamps will overlap a given grid cell on the same day. In these cases, we discard one timestamp, to avoid inconsistent values between grid cells. That is, when accumulating the most frequent label and aggregating values on the overlapping cell, we use only those patches with a timestamp close to that of the neighboring grid cells. This neighboring selection mitigates the

problem of inconsistent values between neary grid cells due only to timing. Finally, we accumulate the aggregated grid-cell values to create the daily files. Given the MODIS orbital coverage, the complete `OC-Gridcell` set contains around (20 + 22 years) × (365 days/year) × (65 160 grid cells) × 32 B ≈ 29.8 gigabytes.

Table 3: Information provided in AICCA for each 1° × 1° grid cell: a cloud class label computed by RICC and a set of diagnostic quantities obtained by aggregating MODIS data over all patch pixels for that grid cell.

Variables	Description	Values	Type
Cell	(lat, long) for grid cell	2	float32
Timestamp	Time of observation	1	float32
Label	Most frequent class label in grid cell	1	int32
COT_1deg	Mean of pixel values in grid cell	1	float32
CTP_1deg	”	”	”
CER_1deg	”	”	”
CPI_1deg	Most frequent particle phase in grid cell	1	int32

2.4 AICCA Intermediate-Level Data

We also provide CSV format version of AICCA that was supposed to a pre-stage for grid-cell dataset. This “intermediate” level AICCA dataset has both features from Patch-level and Gird-level data: As Table 4 shows the column values of csv data, we provide patch level information as well as latitude and longitude data based on Patch-level data, and then composite them into a daily csv file that is used to construct grid-cell product. The CSV data-format is easily handled with Pandas and we decided to publicly open this data.

Table 4: Information provided in AICCA for csv format version: a cloud class label computed by RICC and a set of diagnostic quantities for each patch.

Variables	Description	Type
Label	Label per patch	int64
Timestamp	Time of observation	string
COT_patch	Mean and Std of pixel values in grid cell	float64
CTP_patch	”	”
CER_patch	”	”
CPI_patch	Number of pixels in patch in {clear-sky, liquid, ice, undefined}	int64
Location	latitude and longitude at a center of patch	float64
Platform	Instrument name {Aqua, Terra}	string
Swath	hour and minutes at each patch	int64

Figure 2 42 plots, one per label, of the label's RFO from 2000 February to 2021 December, for each 1° grid cell on a global grid. This application provides a good case study of how AICCA can, by providing data on a familiar latitude-longitude coordinate system, facilitate climate researcher understanding of spatial-temporal cloud patterns and regimes. Reviewing Figure 2, we note first that various of the subfigures display spatial distributions that are associated with patterns of known cloud classes.

3 Download AICCA

3.1 Globus

You need to register Globus https://www.globus.org/data-transfer, a high-speed data transfer service, to download AICCA dataset. AICCA dataset is available at Github: https://github.com/RDCEP/clouds

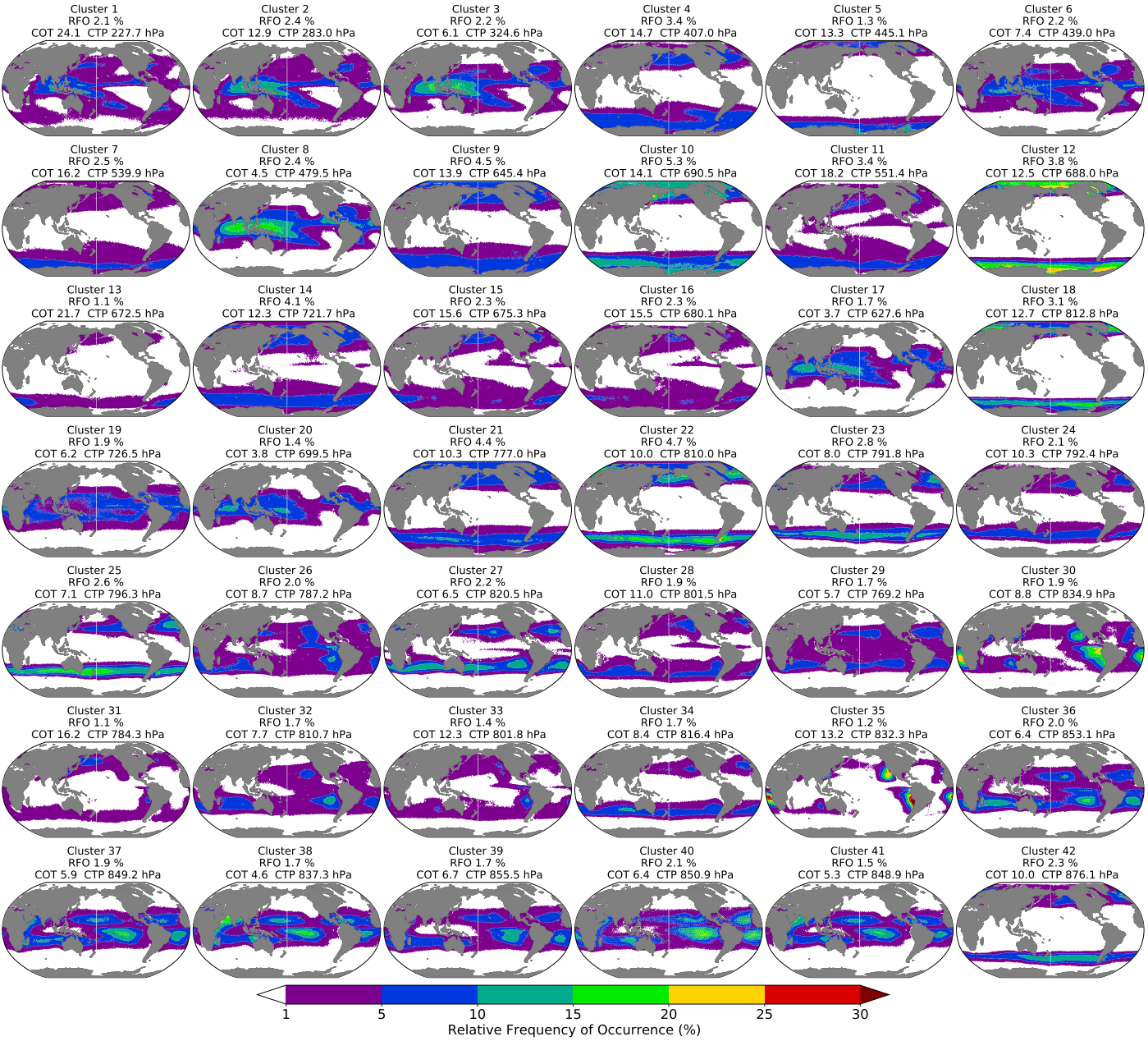


Figure 2: An example application of AICCA. We plot the relative frequency of occurrence (RFO) from 2000 to 2021, for each of the 42 AICCA_{42} clusters. Land is in grey, and areas where RFO $< 1.0\%$ are in white. Surtitles shows global mean RFO, cloud optical thickness (COT), and cloud top pressure (CTP) for the given cluster. We observe that clusters with roughly similar spatial patterns have slightly different global mean COTs, CTPs, and RFO distributions, suggesting meaningful physical distinctions.

3.2 Missing dates

Following dates have no data due to instrument's errors. If you find missing date and/or hour in our data, please post github issue on our github page https://github.com/RDCEP/clouds/issues

Following dates are know for missing dates: **TERRA**:

1) Year 2000 Days: Start from 055, 117, 118, 219-230 has no data.

2) Year 2001 Days: Start from 167-183 has no data.

3) Year 2002 Days: Start from 079-086, 105 has no data.

4) Year 2003 Dys: 351-357 has no data.

5) Year 2008 Days: 356 and 357 has no data.

6) Year 2016 Days: 050-058, and 266 has no data. 266 need check.

7) Year 2017 Days: 078, 144, 168, 260, 351, and 352 has no data.

AQUA:

1) Year 2002 Only 108 days.

4 How to read and use AICCA

To be updated

5 Contact

Machine Learning development: Takuya Kurihana (tkurihana@uchicago.edu)

AICCA data creation and version control: Takuya Kurihana, Ziwei Wang

For any problem with the files and source code, please contact **Takuya Kurihana**.

References

[1] Takuya Kurihana, Elisabeth Moyer, Rebecca Willett, Davis Gilton, and Ian Foster. Data-driven cloud clustering via a rotationally invariant autoencoder. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–25, 2021.

[2] MODIS Characterization Support Team. Modis 1km calibrated radiances product, 2017.

[3] MODIS Characterization Support Team. Modis 1km calibrated radiances product, 2017.

[4] Preesan Rakwatin, Wataru Takeuchi, and Yoshifumi Yasuoka. Stripe noise reduction in MODIS data by combining histogram matching with facet filter. *IEEE Transactions on Geoscience and Remote Sensing*, 45(6):1844–1856, 2007.

[5] Russ Rew and Glenn Davis. Netcdf: an interface for scientific data access. *IEEE computer graphics and applications*, 10(4):76–82, 1990.