Discovering New Global Climate Patterns: Curating a 21-Year High Temporal (Hourly) and Spatial (40km) Resolution Reanalysis Dataset

Chung-Yi Hou*, Robert Dattore, Grace Peng****

*Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, Champaign, IL, United States. <hou@illinois.edu>

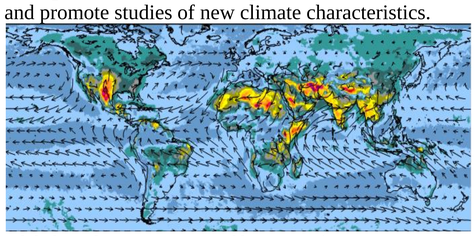
**Computational & Information Systems Laboratory – Data Support Section, National Center for Atmospheric Research, Boulder, CO, United States. <dattore@ucar.edu>, <grace@ucar.edu>

Introduction

The National Center for Atmospheric Research (NCAR) Global Climate Four-Dimensional Data Assimilation (CFDDA) Hourly 40km Reanalysis dataset is a dynamically downscaled dataset with high temporal and spatial resolution that was created using NCAR's CFDDA system.

The dataset contains three-dimensional hourly analyses in netCDF format for the global atmospheric state from 1985 to 2005 (a total of 184,080 files) on a 40km horizontal grid (0.4° grid increment) with 28 vertical levels, providing good representation of local forcing and the diurnal variation of processes in the planetary boundary layer.

Making the dataset publicly available, accessible, and usable will provide a significant resource with greater diurnal cycle details to allow and promote studies of new climate characteristics.



Global Wind Pattern Generated by CFDDA[1].

Methods

The project aimed to make the entire dataset available, accessible, and usable and focused on the following three areas of the data curation process:

Verify Data Quality

- **Consistency:** Comparison of file content and structure between the first data file and the last data file.
- **Validation:** Comparison of the content and descriptions provided by the data files and the dataset's user documentation.
- **Compliance:** Confirmation of data file format and its conformance to Climate and Forecast (CF) Metadata Conventions.

Harvest Metadata Descriptions

- **Tool:** Learning NCAR Computational & Information Systems Laboratory (CISL) Research Data Archive's (RDA) software and processes for automatically gathering metadata from the data files.
- **Format:** Understanding the rationale for CISL RDA's usage of Global Change Master Directory (GCMD) scientific keywords and customized metadata format over standards, such as ISO 19115.
- **Content:** Creating the metadata content through CISL RDA's Metadata Manager.

Document Provenance Information

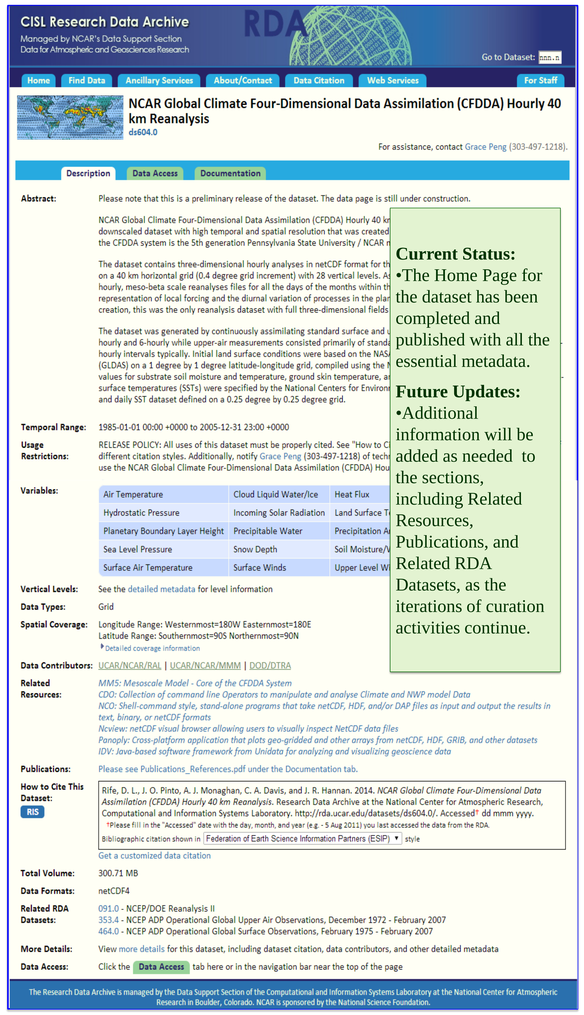
- **The Data Curation Profiles[2]:** Utilize the instructions and worksheets provided by the toolkit to interview the scientists in order to organize and document the provenance relating to the dataset.

Results

The dataset is currently available for access and use at National Center for Atmospheric Research (NCAR) Computational & Information systems Laboratory's (CISL) Research Data Archive (RDA - http://rda.ucar.edu/datasets/ds604.0/).

The dataset's landing page provides three unique tabs with each tab featuring specific information regarding the dataset.

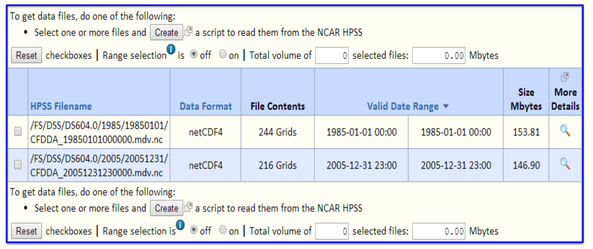
Dataset Home Page
"Description" Tab



Current Status:
- The Home Page for the dataset has been completed and published with all the essential metadata.

Future Updates:
- Additional information will be added as needed to the sections, including Related Resources, Publications, and Related RDA Datasets, as the iterations of curation activities continue.

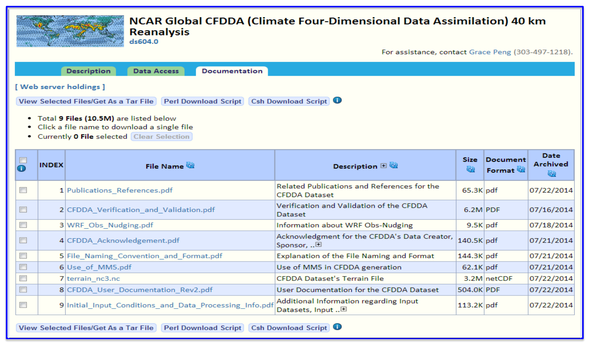
Dataset's Second Page
"Data Access" Tab



Current Status:
- The first and the last file of the dataset have been archived successfully into the RDA.
- Archiving these two files helped test the data ingest and metadata harvest processes.

Future Updates:
- All the remaining data files will be ingested into the RDA and be made available, accessible, and usable.

Dataset's Third Page
"Documentation" Tab



Current Status:
- Information considered to be most critical and helpful to the dataset's usability has been published.

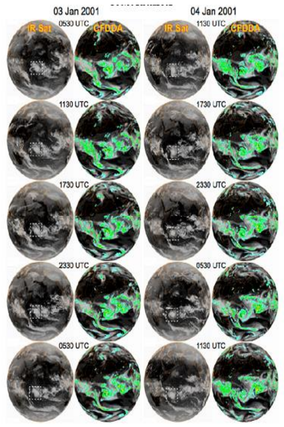
Future Updates:
- Additional documents sharing the dataset team's project-specific knowledge and dataset users' experiences, such as Frequently Asked Questions, will also be made available.

Conclusions

By the time the data curation process started for the dataset, it had been five years since the data files were generated. In addition, at the end of the 18-month project duration, although the PI had generated a user document, the document had not been maintained or updated. Furthermore, not only had the Principal Investigator (PI) moved to a new research institution, but also the remaining team members had been reassigned to other projects. **As a result, the curation process showed that it was crucial to involve data curators during each stage of the data curation life cycle, such as the one proposed by the Digital Curation Center (DCC)[3], instead of at the end.**

Particularly, in the areas of verifying data quality, harvesting metadata descriptions, and documenting provenance information, data curator's skill and knowledge could help the team make decisions, such as file format and structure and workflow documentation, that could have significant, positive impact on the likelihood and the ease of the dataset's management and long term preservation. For example, through the Data Curation Profiles Toolkit's guidelines, the discussions with the Project Manager revealed important information regarding the dataset that was helpful both to promote the data's usability and to enhance preservation planning.

With the proper resources invested in the dataset's curation process, especially with attention paid to the dataset's data quality, metadata description, and provenance information, this dataset is now prepared and can offer much potential to help with new climate pattern discovery.



Time series of regional to global cloud patterns over the eastern hemisphere for 0530 UTC 03 January 2001 to 1130 UTC 05 January 2001. IR satellite imagery from (first and third columns) Meteosat and (second and fourth columns) using the CFDDA dataset is shown. Relative humidity at 500 hPa (gray shades) is shown as a proxy for clouds, and surface rainfall accumulation (colors; green colors denote low intensity, and yellow and orange colors denote high intensity). Dashed white boxes track the visually defined position of tropical cyclone Ando over the Indian Ocean[4].

References

1. National Center for Atmospheric Research, Research Applications Laboratory. (2014). *NCAR Climate Four-Dimensional Data Assimilation (CFDDA) Hourly 40 km Reanalysis User Documentation* (2nd ed.). Boulder, CO: Author.
2. Data Curation Profiles. (2010). *Data Curation Profiles Toolkit*. Retrieved from http://datacurationprofiles.org/
3. Digital Curation Centre. (2014). *DCC Curation Lifecycle Model*. Retrieved from http://www.dcc.ac.uk/resources/curation-lifecycle-model
4. National Center for Atmospheric Research, Research Applications Laboratory. (2010). *Development and Validation of the Joint Effects Model High-Resolution Global Climatography Database*. Boulder, CO: Author.

Acknowledgments

The author would like to thank her mentors, Steve Worley, Terri Betancourt, Andy Monaghan, Daran Rife, Matthew Mayernik, Chi-Fan Shih, Zaihua Ji, and Doug Schuster of National Center for Atmospheric Research (NCAR), Carole Palmer and Cheryl Thompson of Illinois, and the Data Curation Education in Research Centers (DCERC) project, funded by Institute of Museum and Library Services, for providing and supporting this learning opportunity.