Seth Vuletich

LIS 4220

**Data and Metadata Profile**

**Data**

For this project, I have elected to work with to the 4th version of the GISTEMP dataset (<https://data.giss.nasa.gov/gistemp/> ). These data are temperature anomalies for each year from 1880-2020. This dataset is part of a National Atmospheric and Space Administration (NASA) funded project through the Goddard Institute for Space Studies (GISS). Temperature anomalies, in the case of these data, refer to the difference in average monthly temperature from the 30 year average temperature recorded within the years of 1951-1980. The project that produced this dataset was first initiated in the late 1970s by James Hansen. Since then the data have been improved and updated. Nathan Lenssen, Gavin Schmidt, the current project leader, James Hansen, the retired project leader, Matthew Menne, Avraham Persin, Reto Ruedy, and Daniel Zyss are credited with the journal article detailing the 4th version of the dataset. Additionally, Michael Hendrickson, Ken Lo, and Makiko Sato are noted as current participants in GISTEMP analysis. In addition to NASA, GISS, and those referred to in the citation as “GISTEMP Team”, this dataset is highly cited by researchers who should be considered stakeholders. As a federally funded project, the data is within the public domain and, as such, all US taxpayers can be considered stakeholders.

The data from GISTEMP are presented in tables. The core data are divided between 4 files containing average temperatures from 1880 to 2020, 3 of which go into monthly detail and also average the seasonal and annual temperatures for global temperature, Northern hemisphere temperature, and Southern hemisphere temperatures; the 4th file contains annual average temperatures for global zones. There are also 4 derivative tables comparing GISTEMP to temperature data obtained from NASA’s Atmospheric Infrared Sounder (AIRS) program divided into the same categories as the core data. The tables are available in txt and csv formats. While maps are present on the site in netCDF file format as well as Subbox Grid Series binary format (SBBX), the maps are all created from NOAA source data, not the product of GISTEMP analysis. In total there are 8 tables and 9 maps on the site, though only 4 tables are the direct product of the GISTEMP algorithm. While the tables are easily opened with common software like notepad or Excel, the maps are a little more complicated. NetCDF will open in a number of programs including ArcGIS, but SBBX may require conversion using one of the FORTRAN programs available on the site. Lastly, the source code for GISTEMP is available for download, but requires a Unix-like operating system and familiarity with Python to install and use.

**Metadata**

The GISTEMP dataset does not have a metadata file, nor is metadata included within the data files, rather metadata is scattered across the GISTEMP webpage. There is a txt README file containing technical information about the conversion of SBBX map data to other formats, but no such files for the tables or NetCDF map files. Through the various links on the page containing these data, and particularly the FAQs page, one can gain a fairly comprehensive understanding of the data. For instance the source(s) of the data used in the production of GISTEMP are from NOAA’s Global Historical Climatology Network (GHCN) and Extended Reconstructed Sea Surface Temperature (ERSST). From this information I was able to glean that the GISTEMP data are a minimum of three layers of abstraction from the direct temperature measurements. Moreover, these pages discuss the history of the project and the accuracy of the data. There is also a journal article that was released with version 4 of GISTEMP, *Improvements in the GISTEMP Uncertainty Model* (Lennsen et al., 2019), that provides additional details about the methodology used in updating this dataset. Additionally, there are several articles detailing methods used for previous versions. That said, there is no standardization of the metadata and as such, it follows no formal structure.

Likely owing to the non-standardized metadata, finding the GISTEMP dataset was a little challenging. While I accessed it through related resources initially, a simple search for “global temperature rise” in data.gov yielded the sources used to create this dataset, but GISTEMP was elusive, not appearing within the first five pages of results despite being a more accurate fulfillment of the query than many higher ranking results. Even a direct search for GISTEMP brings the actual dataset up 2nd out of 2 results. I think a structured metadata approach would benefit these data enormously. Given the high level of citation for this dataset, I would guess most users reach it through citations and familiarity with its prevalence in the field of climate change, but a direct search is unlikely to find it.

The FAQ section might be more usable if presented as a metadata file and the concept of temperature anomalies explained upfront. It took me at least an hour to familiarize myself enough with the website that I could completely comprehend the data. The map data seem to be far more complicated. There were no clear instructions on necessary programs to access NetCDF files, not to mention the SBBX file type which looks to require some programming knowledge. This is not a problem for common file types, but these specialized file formats are foreign to a novice. Documentation on these file types should be included, and the technical documentation for SBBX files, while important for experts, is pretty inaccessible to the general public, and might waste a non-expert’s time. The single README file on the page details the programmatic requirements to run SBBX files and the naming of it is insufficient to delineate its intended purpose, not to mention the complete exclusion of the data tables produced through GISTEMP.

**Publications**

This is a widely cited dataset, a Google Scholar citation search reveals that the GISTEMP dataset has been cited 231 times, and the current version is cited an impressive 33 times since its release in 2019. The website on which the dataset is hosted lists publications from the team that has worked on this dataset, but no publications from third parties who have also used it. Clearly these data are valuable, but they are presented in a way that is likely to dissuade widespread usage by the general public.

References

GISTEMP Team, 2021: *GISS Surface Temperature Analysis (GISTEMP), version 4*. NASA Goddard Institute for Space Studies. Dataset accessed 2021-01-26 at <https://data.giss.nasa.gov/gistemp/>.

Goddard Institute for Space Studies. 2019. *GISTEMP FAQs*. Retrieved from <https://data.giss.nasa.gov/gistemp/faq/>

Lenssen, N., G. Schmidt, J. Hansen, M. Menne, A. Persin, R. Ruedy, and D. Zyss. 2019. [Improvements in the GISTEMP uncertainty model](https://pubs.giss.nasa.gov/abs/le05800h.html). *J. Geophys. Res. Atmos.*, **124**, no. 12, 6307-6326, doi:10.1029/2018JD029522.