GULF OF MEXICO RESEARCH INITIATIVE
INFORMATION & DATA COOPERATIVE

GRIIDC Dataset Archive and Compression Guidance

Most datasets that are submitted to GRIIDC are organized within a directory structure composed of folders and files. To simplify submission and ensure data integrity, GRIIDC asks data providers to package datasets into a single compressed archive file. The general workflow for packaging files into a single compressed archive file is:

- Put the dataset files and folders inside a top-level folder; name this folder something meaningful
- Make sure none of the files are compressed archives themselves— avoid nesting packages
- Archive and compress the top-level folder and its contents into a single file
- Test the compressed archive file to verify the contents are intact

GRIIDC encourages researchers to package their datasets into either zip (one-step) or tar+gz/bz2 (two-steps) compressed archives. GRIIDC prefers the zip format for most datasets as it is widely used, supports the broadest user base, and has advantages over the two-step tar+gz/bz2 approach. Advantages of using the zip format include:

- Most computers come with the software needed to create or open zip format files
- Archiving and compression processes are combined in a single step, so package contents can be listed and individual files can be extracted, added, deleted, or replaced without having to decompress and then recompress the entire package (as is required in a two-step process)

[Macintosh users note: The built-in archiving utility in the Finder does not properly use the Zip64 extension and the resulting zip files are corrupted if they are larger than 4GB or contain more than 65000 files. Mac users should use Homebrew or MacPorts to download and install free alternative versions of zip format compatible software and run it from the command line instead of the Finder].

However, GRIIDC often finds that files cannot be completely extracted from zip archives when the archive is larger than 100GB. In these cases, GRIIDC recommends a two-step process using the tar utility to pack the dataset into a single file followed by compression using either gz or bz2. Some forms of these utilities (e.g., 7-zip, lbzip2) take advantage of multicore and multithreaded processors and can greatly decrease the time it takes to compress large archives. GRIIDC recommends against creating nested tar files (i.e., tar files that contain tar files) because it requires multiple unpacking steps to extract a file listing or a subset of files. GRIIDC acknowledges this method uses a lot of disk space to process files.

In all cases, GRIIDC requests that researchers test the archives they have created prior to submission as GRIIDC cannot accept datasets that cannot be unpackaged and opened without error.

Summary Recommendations for Archiving and Compression

- Use zip for most files, use tar with gz or bz2 for very large files
- Test archive before submitting to GRIIDC
- Don’t nest multiple archives, as unpacking these requires multiple steps