Application of Advanced Lossy Compression In The NetCDF Ecosystem For CONUS404 Data

Shaomeng Li, Allison Baker, Erin Dougherty, Ryan Cabell, and Lulin Xue | Nvidia and NCAR

Take-home message

Compression is most useful when it preserves both values and conventions.

Motivation

Why necessary yet hard in practice

Large geoscience files are not just arrays. They carry conventions that analysis tools and scientists rely on, in their own ecosystem.

CONUS404 Data

40 years
CONtiguous United States

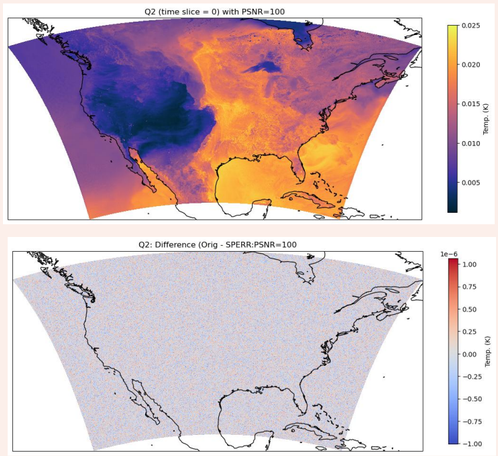
4 km
Resolution

8 PB
NetCDF-4 Files

50k
Users per year

- Lossless compression often gives limited reductions for floating-point model output.
- Advanced lossy compression can reduce size more, but only practical if the workflow remains familiar.
- The practical question: can compressed files still be read, shared, and analyzed just like typical NetCDF files?

Plotting variable **Q2** using SPERR compressed CONOS404. The point-wise difference is also shown.



Method

Bring compression to NetCDF

NetCDF-4 supports "filters" that transform data during write/read, all *transparent* from users.

- **Python / xarray / NetCDF-4 / scripts**

->

- **NetCDF-4 / HDF5**

->

- ***H5Z-SPERR filter***

->

- **NetCDF files, compression Enabled**

Quality Control: pick your target

Bit-per-point | **Peak signal-to-noise Ratio** | **Maximum point-wise error**

Special Sauce: fill value masks

- H5Z-SPERR understands "fill values" and uses a mask to encode them efficiently.
- Special handling of fill value masks further increases compression ration on real values.

transparent to users

Results

Same workflow, compression enabled

Enablement: environment variable

- `export HDF5_PLUGIN_PATH=/my/path`
- `setenv HDF5_PLUGIN_PATH /my/path`
- `os.environ['HDF5_PLUGIN_PATH'] = '/my/path'`

Compression: `nccopy`

```
nccopy -F "VAR0,32028,268651725u,0"
<input_file> <output_file>
```

use a script to run at a scale

Size comparison

Original: 137GB

Lossless: 49GB

Lossy: 14.6GB

(Data of different compression levels is saved to support tasks from quick visualization to high-fidelity analysis.)

What scientists say

- "Easy user experience"
- "Pretty straightforward"
- "Sometimes even faster to plot"
- "Need to be careful about software versions [NetCDF, HDF5, SPERR, etc.]"

Takeaways

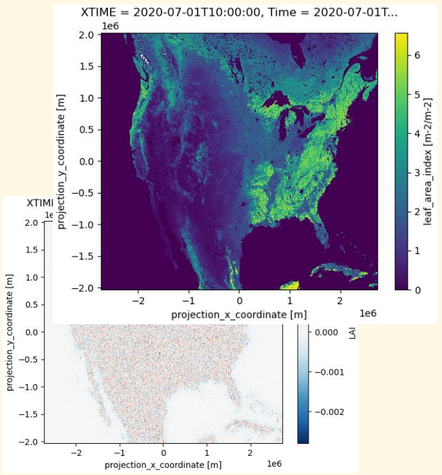
Lossy compression is easy and useful

Core discovery

H5Z-SPERR makes advanced lossy compression behave like a practical NetCDF/HDF5 capability for real geoscience data. It brings the benefit of lossy compression to where it is needed.

- Fill-value preservation is required for trustworthy real-world datasets.
- CONUS404 is a useful stress test because it is large, familiar, and convention-rich.

Plotting variable **LAI** using SPERR compressed CONOS404. The ocean is covered by a fill value mask, which is correctly preserved, resulting in absolutely no difference in the difference plot.



Useful Links:

- SPERR Compressor: github.com/NCAR/SPERR
- HDF5 Plugin: github.com/NCAR/H5Z-SPERR
- Python package ***hdf5plugin***: github.com/silx-kit/hdf5plugin
- CONUS404: gdex.ucar.edu/datasets/d559000