

Data Structure and Format

For attribution, please cite this work as:

Cal-Adapt. (2026). *Data Structure and Format*. Cal-Adapt.
<https://analytics.cal-adapt.org/data-tools/data-documentation/data-structure-and-format.html>

The Analytics Engine public AWS S3 bucket named `cadcat`, short for Cal-Adapt: Data Catalog, contains dynamically downscaled WRF data in Zarr and statistically downscaled LOCA2-hybrid data in both NetCDF and Zarr formats, along with other datasets such as station data (HadISD). More detailed information about the climate data used by the Analytics Engine can be found in the [Climate Model Simulations](#) section.

NetCDF

The raw LOCA2-Hybrid datasets are stored as NetCDF files in S3 at the subdirectory [LOCA2/aaa-ca-hybrid](#). The advantage of NetCDF files is that they are single files stored in a directory structure, and can thus be manually downloaded using a web browser for a minimal set of data. The disadvantage of NetCDF is that they need to be loaded completely into memory to access even a slice of the data, which can be resource exhaustive or even prohibitive.

Zarr Stores

The majority of the climate data for the Cal-Adapt: Analytics Engine stored on AWS S3 is in a cloud-optimized multi-dimensional array format known as Zarr. Zarr stores can be randomly accessed to retrieve a slice of the data and do not require that the entire dataset be loaded into memory. This is vitally important when dealing with hourly data over large spatial and temporal extents. The Zarrs are stored in S3 in a hierarchical directory structure designed to work in conjunction with Intake ESM to make the data discoverable, searchable, and easier to access. The directory structure is prescribed by the [intake ESM](#) system and can be non-intuitive to users not familiar with it.

Zarr stores have many directory levels, which make it difficult to download manually using a web browser. Either the use of AWS's command line interface (AWS CLI) for S3 or Python programming is needed to access and/or download the data in this format. A web browser can be helpful in understanding the data structure before attempting to access specific files. AWS allows users to browse S3 buckets via a web browser - here is the top level of the `cadcat` bucket [AWS S3 Explorer](#).

There are 4 main collections of Zarr datasets in the S3 bucket:

1. **WRF**: Dynamically-downscaled global climate models performed by the Weather Research and Forecasting model
2. **LOCA2-Hybrid**: Statistically-downscaled global climate models performed by LOCA (Localized Constructed Analogs) version 2 for California
3. **HadISD**: Hadley Centre observations datasets (station data). Does not conform to the Intake ESM data structure, and is not part of the Intake ESM data catalog.
4. **HistWxStns**: Historical observation datasets (station data), including the HadISD stations.

WRF and LOCA2-Hybrid follow similar directory structures with some notable differences:



- LOCA2-Hybrid has an additional level (Variant/member_id)
- Not all parameters are the same between WRF and LOCA2-Hybrid
- Parameters are not always internally consistent.

More advanced query tools are useful in figuring out where the data is and how to load it into different software packages. The Analytics Engine searchable [data catalog](#) was generated from the contents of the Intake ESM catalog to help identify the exact path to the data of interest. Users can type into the search box to identify datasets of interest and view the path to those datasets in the S3 bucket structure.

