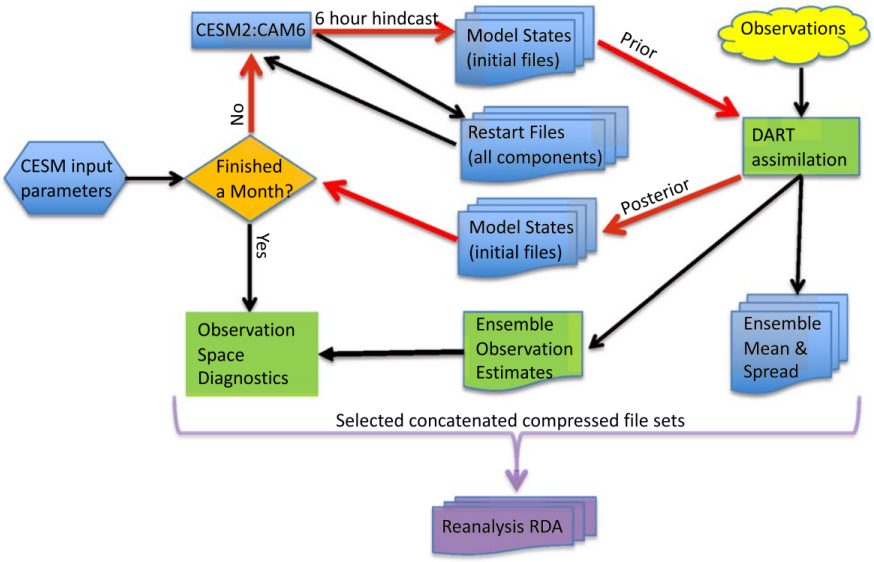
CAM6 Data Assimilation Research Testbed (DART) Reanalysis Use Case

Compiled May 2022 by Model Data RCN team

Summary
Weighted rubric score - 59
Category - Preserve selected simulation workflow outputs

- Use Case Description
    - High-level overview of the use case
        - This project is an atmospheric data assimilation reanalysis for Earth surface model forcing.
        - Ensemble data assimilation - simulating possible atmospheres, guided by observations. Provides an estimate of uncertainty.
        - Need to maintain a certain spread in the forcing data. This is important for maintaining spread in the surface model data assimilation, for diagnostics, and for potentially comparing with other reanalyses.
        - Simplifies the forcing for others who are running Earth surface models
        - Published journal link:
            - https://doi.org/10.1038/s41598-021-92927-0
        - NCAR data repository link:
            - https://doi.org/10.5065/JG1E-8525
        - Repository description:
            - “These CAM6 Data Assimilation Research Testbed (DART) Reanalysis data products are designed to facilitate a broad variety of research using NCAR's CESM2 models, ranging from
                - model evaluation to (ensemble) hindcasting,
                - data assimilation experiments,
                - and sensitivity studies.
            - They come from an 80 member ensemble reanalysis of the global troposphere and stratosphere using DART and CAM6 from CESM2.1. The data products represent the actual states of the atmosphere during the most recent decade at a 1 degree horizontal resolution and 6 hourly frequency. Each ensemble member is an equally likely description of the atmosphere, and is also consistent with the dynamics and physics of CAM6. Visit DART Reanalysis for details.”
            - “This large dataset (~ 120 Tb) has a unique combination of a large ensemble, high frequency, a global model constrained by observations, and multiyear time span, which provides opportunities for robust statistical analysis and use as a machine learning training dataset.”

        - "The data differs from model generated training sets in that it is constrained to represent the atmosphere (and plant growth characteristics), rather than just the model formulation."
    - Science goals and basic workflow
        - Run a month of assimilation at a time. Got a huge amount of output, discarded a lot along the way due to disk space limitations. Grouped data into larger files
        - Do a short forecast using all of the ensemble members, then add in the data assimilation.



        -

- What use-case specific additional materials were preserved and shared?
    - Data
        - 120 TB of data divided into 6 categories of products. The NetCDF files are on the CAM6 grid. The observations files data are at the observation locations and times (ungridded).
            - **CESM Flux Coupler (cpl7) Files: 14.16 TB.** These are used to "force" (provide the upper boundary conditions to) the surface models.
            - **CESM Atmosphere (CAM6.0) Files: 4.68 TB**
                - "The full ensemble ("allinst") of model states from the end of the 6 hour hindcasts ("forecast"), possibly inflated by DART's adaptive inflation algorithm ("preassim")." Frequency is weekly.
            - **External System Processing (DART) Files: 13.22 TB**
                - Observations files contain the observations assimilated by DART, DART's ensemble estimates of the observations, and quality control data which can be used as the "labels"

in machine learning. They are ungridded. The data is available every 6 hours within month-long files.
        - Ensemble means and standard deviations of both the “forecast” (or “preassim”) and “output” (analysis) stages every 6 hours within month-long files.
        - “Observation space” diagnostics; pictures of the comparison of the estimates of the observations to the actual observations.
    - **CESM Land Model (CLM5.0) Files: 3.05 TB**
        - CLM history files, each containing a year of plant growth variables from 1 member (INST)
    - **CESM Restart Files including Initial Files: 84.62 TB**
        - All the file types required to (re)start a single instance CAM6 hindcast, with 80 instances available.
            - “They can be used to start (ensemble) hindcast experiments and further CAM assimilation experiments.”
    - **Input Files Not Available Elsewhere: 280.86 GB**
        - Observations files, as above, but without DART’s estimates and quality control data.
        - The SST data which provided some of the lower boundary conditions for CAM6.
- Inputs to model
    - Total data volume preserved in a repository by the PI
        - None
    - observation sequence files - archived all of these, have original observations and all model estimates of the observations. Also have quality control flags that indicate if the observations were used. These can be used as labeled data in machine learning applications. This is a potential use of the data.
    - Spin up files are not publicly available, but are on the NCAR Campaign store. These are usually not available to others because they aren’t useful. The final state is provided in the archive because this is the starting point for the reanalysis.
- Raw model output
    - Flux coupler and restart files (could also be used as input for others)
    - ensemble members - this is not often provided by reanalyses, they are often only provided as ensemble means. But these are included in this data
    - CLM history files
    - Data that was not included were products that were necessary to run and evaluate the model output, but once they were used they were removed because they were not relevant for the product

            - Ex. restart files were saved once per week, but 28 were generated. So the other 27 were removed, unless a problem occurred during the modeling.
            - Ex. log files. Many were generated but were repetitive and not needed after the runs were completed.
        - Total data volume not preserved in a repository? (might be retained on PI's local working storage)
            - The spinup assimilation is archived in Campaign Storage and is 220 Gb.
    - Processed model output
        - Also do post processing - every month do a comparison of the reanalysis against the observations. Look for biases, where it may be breaking down. See "Observation space", above.
        - Adding data on an ongoing basis when the reanalysis is done.
        - Exported a small slice to the AWS cloud, specifically for people who were interested in the atmospheric forcings on plants.
        - Total data volume not preserved in a repository? (might be retained on PI's local working storage)
            - None
- Software
    - Model configuration
        - CESM: cpl, cam, clm, cice, mozart, esp (DART)
        - Flux coupler (cpl7) used to drive model inputs/sequencing.
        - Configuration is embedded within the DART and CESM codes that are archived on the Github sites. There is a third Github repository that describes the specific case, sort of a diary of the experiment.
    - Preprocessing code
        - The code used to translate the original Sea Surface Temperature files (AVHRR at ¼ degree resolution and daily frequency) into the form required by CESM have been archived on Github and the data files are provided.
    - Model code
        - https://github.com/kdraeder/cesm
        - There are minor changes to the CESM scripting to make it run more efficiently in ensemble mode. These are available in https://github.com/kdraeder/cime; in the cime_reanalysis_2019 branch.
        - CAM is publicly available, managed by others. The version used here was a slightly modified form, available via github
        - Data assimilation code (DART) also available publicly via Github. The Reanalysis used the "reanalysis" branch.
    - Postprocessing code
        - Also available in DART code. For this project, they used Matlab to do the postprocessing.

    - Other
        - Documentation
            - The published paper linked at the top is an important source of documentation
            - Github Link:
                - https://github.com/NCAR/DART/wiki/1-degree,-CAM6,-ensemble-reanalysis-for-CESM-experiments-(2011-thru-2019):-DATM,-hindcasts,-model-evaluation#cesm-flux-coupler-cpl7-files
                    - This is a summary of the project.
                - There is another github repository about this case specifically, providing more details about the project. Select the "f.e21.FHIST_BGC.f09_025.CAM6assim.011" branch.
        - Visualizations or images
            - Within the journal article related to this dataset, workflow visualizations are present.
            - Post processing is mostly images, examples shown in Scientific Reports articles. People can look at these. A complete set is available in the archived data (see "Processed model output", above). Some other visualizations are on a web site that is not easily accessible, so may not be maintained for long. They are year to year comparison for data, e.g. to see changes in a given month over time.
- Why were these things preserved and shared?
    - General
        - From the published paper
            - "The first motivation for the creation of this dataset was to provide that forcing in the context of an Earth system modeling framework, in which the atmospheric forcing can be applied to the non atmospheric components consistently and conveniently."
            - "Such a dataset is challenging to create and archive, in terms of both computer resources and personnel time, as well as requiring careful consideration of the model definition and DA tuning. We believe that making it freely available will accelerate research in the Earth sciences."
        - To support people doing research with surface components of CESM who need forcing data for their experiments.
        - By looking at certain combinations of the data, it is possible to study model biases, which helps in the development of CESM / CAM.
    - Reasons why the things listed above are important
        - Expected/intended audience and what they expect/need
            - Are there specific people who will be using the data downstream?

            - Yes, specific users are interested in this data. People who want to do data assimilation using the land, ocean, or sea ice models as their forecast model will benefit from using this forcing data.
        - Possible/aspirational users?
            - Machine learning applications.
    - Note any temporal considerations, such as particular products that become more/less useful over time
        - Will be making updates going forward.
        - This will serve as a record of the weather over the years covered by the project. So will be useful for anybody studying this time period
        - They did a previous study of 2000-2010 with a previous version, and people are still using it. Hope people will transition to the new version.
        - This took 18 million core hours. These resources are not available to everyone.
- Broader Impacts:
    - How will output from this project be used by stakeholders?
        - There are many use cases. E.g. Improving models by identifying biases.
    - How were stakeholders involved in the data curation decision-making?
        - They queried the community when they were beginning the project, and asked if there were any data that people would want to have archived. The plant modelers asked for some plant fields, which is why these were retained and made available separately via AWS.
    - How will stakeholders be compensated for their participation in the data curation decision-making process?
        - Free access to all of the data, and support from the data creators in understanding and using it.
- Do you have any concerns about misuse of your data or software? If so, what concerns do you have, and what are the reasons for those concerns?
    - No worries. It is all public property. It is esoteric enough that it is hard to imagine people misusing it for political reasons. People theoretically could try to use it to attack the project, but it is part of the scientific process for people to examine the work to find problems.