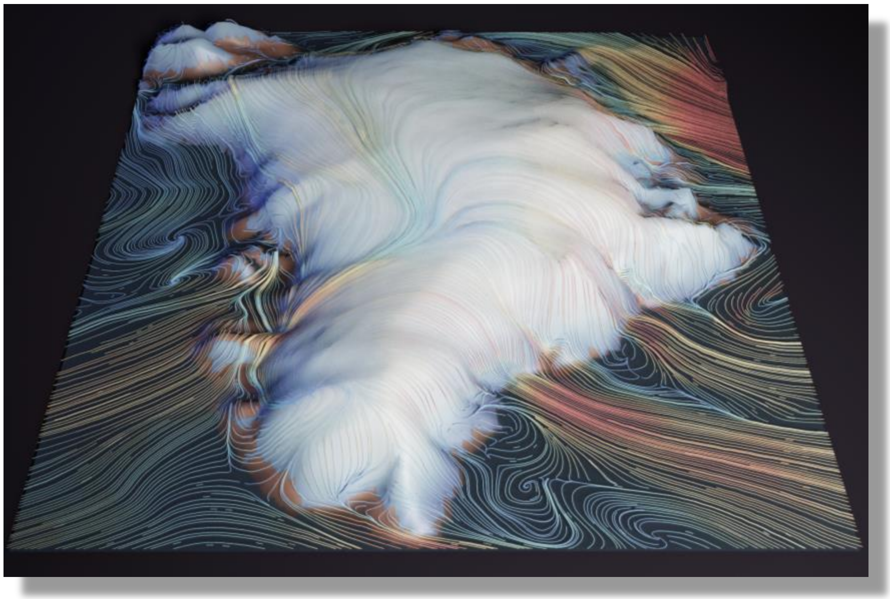
CESM

Community Earth System Model



Data Management and Data Distribution Plan for the CESM Project

January 2022

Table of Contents

Cover image: Snapshot of the lowest model level streamlines, draped over the Greenland ice-sheet and colored by wind speed. Simulation was performed with a 1/8° refined grid over the island of Greenland using the variable-resolution configuration of the spectral-element atmospheric dynamical core in CESM2. Katabatic winds can be seen accelerating down the eastern slopes of the ice sheet. Visualization was developed by Matt Rehme (CISL) and Adam Herrington (CGD) of the National Center for Atmospheric Research, and was inspired by a visualization of winds over Antarctica by the Polar Meteorology Group at the Byrd Polar & Climate Research Center.

Acknowledgments

The CESM Project is supported primarily by the National Science Foundation (NSF). The National Center for Atmospheric Research (NCAR) is a major facility sponsored by the NSF under Cooperative Agreement 1852977. Computing and data storage resources for the CESM Project are provided by the Computational and Information Systems Laboratory (CISL) at NCAR. We thank all the scientists, software engineers, and administrators who contribute to the CESM project. This Data Management and Data Distribution Plan was updated and edited by Gokhan Danabasoglu (the CESM Chief Scientist) with input from the CESM Working Group Chairs and Liaisons and the members of the CESM Scientific Steering Committee. Contributions from Gary Strand, Adam Phillips, Dave Hart, and Jeff de La Beaujardiere are particularly acknowledged.

Introduction

The Data Management and Data Distribution Plan for the Community Earth System Model (CESM) Project documents the procedures for the storage and distribution of data produced by the CESM Project on the National Center for Atmospheric Research (NCAR) computers with the CESM allocation. These procedures reflect the approaches, standards, and conventions that coordinate the production, post-processing, distribution, and storage of simulation data agreed upon by the CESM Scientific Steering Committee (SSC) and the CESM Working Groups (WGs) that comprise the CESM Project. The overall goal of this plan is to provide the best possible access and ease-of-use of CESM Project data to a broad and diverse user community within the constraints of available resources, following FAIR (Findable, Accessible, Interoperable, and Reusable) data principles.

Users of CESM Project data span a wide range of interests that include: Scientists and researchers at universities, federal laboratories, and NCAR; CESM WGs performing the simulations; impact analysts; national and international assessment programs, including those of the U.S. Global Change Research Program (USGCRP) and the Intergovernmental Panel on Climate Change (IPCC); Coupled Model Intercomparison Project (CMIP); policymakers; regional climate modelers and other modeling groups using CESM Project data as forcing input to their models; and industry. The broad common needs of these users include production of useful datasets; easy and ready access to the data; diagnostics of CESM scientific and computational performance; and various types of analysis and related tools.

In accordance with the National Science Foundation (NSF) data policy and the NCAR mission, the CESM Project is committed to the timely availability of results from CESM model runs for publication and sharing of the scientific data generated by CESM Project research activities. Open access to CESM data products is essential to the Project. Analysis and interpretation by the broader community promotes scientific discovery, and it leads to new insights into model behavior that feedback into model development efforts. At the same time, however, the efforts of CESM developers and the designers of the scientific simulations performed with the model under the auspices of the CESM WGs need to be recognized by providing them a reasonable amount of time for first access to the simulations. It is the intent of this Plan to define the guidelines to meet these objectives. These procedures thus apply to all CESM data created under the auspices of the CESM WGs, using the NCAR computational and storage resources allocated to the CESM Project.

CESM simulations use lossless compression by default to reduce their data volume. An evaluation of various levels of lossy compression is also in progress. To help with further reductions in data volume and to provide guidance to the community in their choices of output fields, the CESM Project will soon offer several sets and levels of default output choices that can be activated by a CESM-wide setting. A default setting of a low-level model output that includes basic fields for model evaluation and analysis is anticipated to reduce stress on storage, particularly for the university users.

This plan supersedes all previous CESM Data Management and Data Distribution Plans. The procedures below are not intended to be retroactive to currently available data. Recognizing that climate modeling is an evolving field, the CESM SSC reserves the right to update the guidelines and procedures outlined in this document. Any changes will consider resource implications for the CESM Project. When changes in data policy require substantial increases in equipment, supplies, or personnel resources, pre-existing datasets will not be necessarily expected to comply with these changes.

Scope

This Data Management and Distribution Plan documents the procedures for the release, distribution, and storage of data produced by the CESM Project on NCAR computers with the CESM allocation. CESM data storage allocation / resources on NCAR infrastructure does not cover the storage of data from CESM simulations carried out: in partnership with other national and international collaborators; using Accelerated Science and Discovery (ASD), NCAR Strategic Capability (NSC), or NCAR Laboratory allocations; or other projects funded by various agencies. This document also provides suggested pathways for the storage of data from these latter efforts. Regardless of how they are produced, all CESM datasets are encouraged and expected to follow the naming, distribution, etc. guidelines provided here.

Data Manager

Many of the procedures described in this document are managed and enforced by the CESM Data Manager. Specifically, the CESM Data Manager is responsible for management of data generated by the CESM Project and used by its community; efficient, effective, and innovative distribution, access, and use of CESM data; conducting requirements analysis for data management policies, plans, high-level design, and implementation, including software and hardware to meet these requirements; designing, coding, and documenting programs that ingest, reduce (via compression), analyze, and visualize the data; archiving and cataloging these datasets and any derivative datasets created by these processes; and providing metrics and analyses of CESM data storage costs, accesses, and other relevant measures to stakeholders to the extent possible. The Data Manager serves as the primary point-of-contact and consultant for the CESM Project and the community on data-related issues.

Data Categories

For the purposes of this plan, CESM data are broadly categorized into Development and Community simulations, primarily for consistency with the designations used for managing the CESM allocation. The main defining characteristic of Community simulations is that the majority of such simulations have broad appeal across the climate science community and are thus made available for community access and analysis. Examples include simulations that contribute directly to coordinated national or international modeling activities, e.g., CMIP, and benchmark simulations that document CESM components and new coupled configurations and capabilities of the model (see below). Both are examples of simulations where the project benefits directly from analysis and interpretation by the broader research community. The Community simulations also include experiments that are designed and conducted by a subset of WG members to understand, for example, certain model behavior or phenomena.

The outcome of community analysis can lead to development efforts, such that development and assessment / evaluation activities are synergistic and sometimes become blurred. In contrast to Community simulations, however, Development simulations may not necessarily be made available for analysis beyond the WG members who produce the runs. They include, for instance: simulations to understand fully-coupled CESM or a component model behavior; document biases and determine the responsible processes; efforts to improve the representation of processes; and activities to add new capabilities to CESM important for advancing simulation fidelity, for new community-based science and for future releases of the model.

Within these two general categories, CESM simulations can be further grouped as follows:

CESM Development Simulations

CESM Testing integrations: CESM Testing integrations are very short runs (typically model days to years) carried out to cover three traditional and well-defined tasks: verification of model functionality; performance tuning; and debugging.

CESM Evaluation integrations: CESM Evaluation integrations are relatively short runs (typically model days to several decades) to examine specific model behavior, such as the response to changes in the representation of physical processes or boundary conditions, or to validate a port of the code to a new computing platform.

CESM Community Simulations

CESM Control integrations: Control integrations are typically hundreds to thousands of model years in length, with constant forcing over time, and use a code base that corresponds to a major CESM community release or tag. Control integrations define the basic long-term climate of CESM. A Control run needs to be long enough in simulated time for slow climate adjustment processes, such as subsurface water in the land model, to come into some reasonable balance. Some processes, such as the deep ocean heat and salinity, may take thousands of years to reach an equilibrated state. The standard data output usually corresponds to monthly averages, with perhaps daily or higher frequency output over subsets of the Control and perhaps for fewer, select variables. Higher frequency data output for special analyses or to drive regional-scale models may also be provided from Control simulations upon specific requests or to meet requirements of nationally- or internationally-coordinated experimental protocols.

CESM Experiment integrations: CESM Experiments are typically tens to hundreds of model years in length (or even much longer; for instance, several millennia in paleoclimate applications) and are usually made with modifications introduced into the control version of the CESM and its components to conduct a specific scientific experiment or examine a policy scenario. A CESM Experiment simulation may be a single run or a series of runs. The modifications may be either in the representation of processes in the model, the forcing data, or both. Often, a given experiment design will include the production of an ensemble of runs to provide an estimate of the range and significance of the model response to the imposed modifications from a control case. In this circumstance, the ensemble of runs constitutes the data for the particular Experiment. The Experiment category also includes simulations that are designed and conducted by a subset of a WG's members to understand, for example, certain model behavior or phenomena. These latter simulations clearly serve the broader community by helping advance our understanding of the Earth system.

Summary of Simulations

These different categories of integrations produce output data of various forms as defined in Appendix A. In the procedures below, the different categories of runs identified above will be designated simply as *Testing*, *Evaluation*, *Control*, and *Experiment*.

Ownership Rights and Responsibilities for Data

CESM output data created with the CESM Project resources are considered to be owned by either the SSC or the CESM WG(s) that designed and performed the runs. The Chair of the SSC (the CESM Chief Scientist) and the Co-chairs of the WG(s) are designated as the Principal Investigators (PIs) for the respective simulations, acting on behalf of the SSC and the WG(s), respectively. In some instances, the lead of a subproject within a WG can serve as a PI for the related set of Experiment simulations. The PIs are responsible for ensuring the simulation data have been checked for quality control (see Data Quality Assurance section), have been appropriately documented, and are made accessible, as appropriate, following the data release procedures documented below.

Testing: The data from CESM Testing integrations are the property of the PIs of the relevant WG(s) and are considered to be non-public and internal. CESM Testing integrations are only required to be documented to the extent needed by the PIs or WG members.

Evaluation: CESM Evaluation integrations are treated in the same manner as CESM Testing runs.

Control: These data are the property of the SSC, with the CESM Chief Scientist serving as the PI. CESM Control integrations are documented on the CESM Experiments and Output Data Website (http://www.cesm.ucar.edu/experiments/). This includes descriptions of the simulations and pointers to the validation plots and data files.

Experiment: These data are initially the property of the CESM SSC, WG, or WGs that design and conduct the Experiment, with the CESM Chief Scientist and / or the Co-chairs of the WG(s) serving as the PIs. Examples of CESM Experiment simulations that fall under the purview of the SSC include those performed in support of either national or international assessment activities or other coordinated model protocols. Most other examples of CESM Experiments fall under the purview of one or more WGs. Transfer of access to a broader community of these CESM Experiment simulations takes place over time as defined by the data release procedures below. Like Control integrations, CESM Experiment integrations are documented on the CESM Experiments and Output Data Website.

When multiple WGs collaborate on Community integrations, the Co-chairs of those WGs serve as PIs, although it is acceptable for the cochairs to designate only one WG or other related PIs as the primary owner of the data.

Data Release Timeline

The intellectual investment and time committed to the design and execution of a CESM Control or Experiment simulation entitles the PIs, acting on behalf of the SSC and the WGs, to the first benefits obtained from the resulting data. Publication of descriptive or interpretive results derived immediately and directly from the simulation data is the privilege and responsibility of the PIs. However, to further CESM science objectives, the PIs are encouraged to share their data with colleagues prior to the release deadlines. When the release deadlines for CESM Project data are reached, the data move into the public domain[1]. In special cases, the CESM Data Release Timeline can be superseded by the SSC. Examples of this could be the involvement of the CESM Project in national or international modeling or assessment activities, or other coordinated model experiment protocols, where the timely release of CESM data may be deemed necessary.

The release status of CESM data is characterized as being Protected, CESM Access, or Public. Protected data are owned by the PIs, as defined above. CESM Access data are Protected data that have been made available to all CESM WG members. Protected data become CESM Access data, then Public data, by permission of the PIs or through the expiration of the proprietary time period as defined below. Public data are open to access by the broader community.

[1] The CESM data, including those submitted to the CMIP6, are licensed under a Creative Commons Attribution – ShareAlike 4.0 International License.

Procedure

All CESM data are initially designated as Protected.

All Evaluation and Testing data remain as Protected unless the PIs decide otherwise.

CESM Control data become Public once the Control integration has been completed and validated by the CESM Chief Scientist.

CESM Experiment data shall be available to members of any CESM WG (Access data) *no later* than six months following the conclusion of the full set of Experiment integration.

CESM Experiment data shall become Public as soon as a scientific paper on the results has been submitted by the PIs or one year after the end of the simulation, whichever comes sooner. If a scientific paper is written by a member of a CESM WG after the data became designated as Access data, but before the one-year deadline or a paper is submitted by the PIs, the PIs are encouraged to make the data Public.

Any person wishing to make use of Experiment data before these dates should communicate directly with the PIs about access to the data. In this circumstance, it is expected that the PIs will be offered co-authorship of any published results, which they can accept if they wish.

In special cases, when more timely delivery of CESM data is necessary and / or is in the best interest of the CESM Project, the CESM Data Release Timeline can be superseded by a decision of the SSC.

Information regarding CESM Public data availability for various simulations, including appropriate references and acknowledgments, will be made available on the CESM website. The users of these data are fully expected to include the acknowledgments provided on the website in their publications (https://www.cesm.ucar.edu/publications/#acknowledge).

Data Retention

The key to management of CESM data is to have data from Development and Community simulations stored and distributed via different strategies, with each tailored to suit the needs of the CESM user community. **The guiding principle is that storage capacity is limited, and it will NOT grow as fast as needed to retain all CESM output indefinitely.** The CESM data retention procedure attempts to strike a balance between the scientific need to retain data from older simulations with the growing cost of doing so in a resource-limited environment. Unlike observational data, model simulation data often become less valuable with time as better models of higher quality are developed and integrated. Nevertheless, publication of scientific analyses of CESM Control and Experiment integrations can continue for many years after the data are generated.

Accordingly, data from CESM Control and Experiment integrations shall be preserved for specified time periods to allow extraction of the scientific content.

CESM Project data at NCAR will be retained under the guidelines of this data stewardship plan. The CESM Data Manager, in consultation with the PIs of the data, is responsible for the stewardship. The datasets created by the CESM Project at other sites in collaboration with its partners are expected to follow the same guidelines. Sites holding CESM Control or Experiment integration data that are Public should give the CESM SSC the option of archiving these data at NCAR before deletion, subject to availability of storage resources at NCAR.

This section details retention policies for CESM *history* output, while Appendix A presents retention policies for other types of CESM data.

Procedure

Testing and Evaluation data: Output data will be stored at NCAR for a maximum of 2 years after creation, at which point they will be removed, unless the relevant WG cochairs make a request to extend the retention for cause to the CESM Chief Scientist and the CESM Data Manager. Where appropriate, earlier removal of these data is strongly encouraged. Should storage resources become an issue, the SSC reserves the right to intervene upon the recommendation of the CESM Data Manager.

Control and Experiment data: Output data will be stored at NCAR for a period of four years. These data will then be gradually reduced to 50% of their initial volume over a period of two additional years, based on usage and anticipated demand. This data level will be maintained for two more years. Afterwards, the CESM Data Manager and each CESM WG / PI will determine data to be removed and at what rate, as the archived data are gradually reduced to an acceptable level, also taking into account allocations. In all stages, while data volume reductions will be accomplished by a combination of lossy compression and selective deletion of unused data, labor cost of this effort will be weighed against the storage costs at the time. Datasets designated as high-use / -value, such as those from CESM CMIP simulations or various Large Ensembles, may have retention periods longer than 8 years. Exceptions for longer retention periods must be approved by the CESM Chief Scientist or the CESM SSC.

The CESM Data Manager will be responsible for making sure the above procedures are implemented and followed by the PIs.

Output Data Format and Metadata

Standard data and metadata formats are crucial for the automated analysis necessary to efficiently interact with large data collections (see Appendix B). CESM uses netCDF as the standard data format for all output data. The use of netCDF makes CESM output data readily accessible to a variety of existing graphics and analysis packages.

In the broadest sense, metadata are simply “structured data about data”, describing important attributes of an information resource. Metadata for CESM data are carried in the header section of the model output netCDF files. CESM data will comply with the metadata requirements or standards (e.g., ISO 19115) of data repositories to the extent possible.

Procedure

All CESM components will create netCDF output history data.

All post-processed CESM data will be made available in netCDF format.

All CESM netCDF output history data will comply, to the extent possible, with the Climate and Forecast (CF) metadata conventions.

Output Filename Conventions

The CESM Project has adopted naming conventions for output files. CESM output files fall into two broad categories: i) those generated by the CESM component models at run-time, i.e., model output data; and ii) those created by post-processing of the run-time files, i.e., post-processed data. The naming conventions are described on the CESM website at https://www.cesm.ucar.edu/models/cesm2/naming_conventions.html. All CESM simulations must conform to these output filename conventions.

Data Quality Assurance

Quality control of CESM data is carried out by relevant WGs. Primary responsibility for quality control of CESM data products lies with the PIs overseeing the model integrations. The PIs should address questions raised by the researchers using these data as quickly as possible. If an issue with the data is found, this should be noted on the related project's website, and a notice should be sent to relevant mailing lists. Depending upon the severity of the issue, the data may be withdrawn from all outlets, e.g., spinning disk, the Climate Data Gateway (CDG), and Earth System Grid Federation (ESGF) nodes. For data that needs to be withdrawn from the ESGF nodes, a formal description of the issue should be filed with the ES-DOC errata website, and the relevant Computational and Information Systems Laboratory (CISL) ESGF contact should be notified for the data to be pulled.

Access to Data

Web technologies allow for efficient discovery and access of CESM data. Currently, the primary distribution system for CESM Project data is the CDG. Other means of access to CESM Project data include the University Corporation for Atmospheric Research (UCAR)

Graphical Information System (GIS) portal; the NCAR Geoscience Data Exchange (GDEX), formerly known as the Digital Asset Services Hub (DASH) Repository; ESGF; repositories at non-NCAR centers; and cloud-based services. (See below for more details on the last two access options.)

Due to privacy concerns and the open and distributed nature of CESM datasets, the CESM Project cannot track statistics on data download and use.

Procedure

In general, output data from CESM Testing and Evaluation integrations will be made available only to the WG members that are directly involved in the development experiments. For WG members who do not have access to the NCAR systems, these data will be made available via the CDG (http://www.earthsystemgrid.org/).

For Control and Experiment integrations, to maximize the ease-of-access and value of the data to the scientific community, all CESM Public data shall be made available via the CDG.

All CESM data available via ESGF are accessible without any restrictions, as required by ESGF policy.

NCAR will serve as much CESM data online as possible. Other centers archiving and serving CESM data are encouraged to do so as well in collaboration and coordination with the CESM Project through the relevant PIs and the CESM Data Manager as indicated below.

Initial and boundary condition datasets are available as part of the model releases or tags. Restart files are also publicly available from the Control and Experiment simulations. However, due to low demand, they may be obtained upon request from the CESM Data Manager.

CESM Data Produced by other NCAR Projects

CESM simulations are also carried out by NCAR researchers using ASD, NSC, and NCAR Laboratory allocations. As indicated earlier, the CESM Project data storage allocation on NCAR infrastructure does *not* cover the storage of data from these other CESM simulations. This document can instead be considered as best practices for preparing and documenting CESM datasets generated at NCAR. In rare cases, exceptions allowing such datasets to be stored and maintained by the CESM Project may be granted by the CESM Chief Scientist – if needed, in consultation with the CESM SSC and CESM Data Manager, for CESM simulation datasets that are deemed to have significant community value, are distinct from available CESM Project datasets, and contribute to the scientific objectives of the CESM Project.

Procedure

The long-term storage needs for other NCAR-generated CESM datasets must be identified and budgeted, as appropriate, in the related proposals according to the NCAR DASH process during proposal preparation. Storage options include making use of Laboratory storage resources and purchasing storage capacity on the CISL infrastructure at the NSF-approved rate for NCAR of $45 / TB / year (with possible additional data processing costs if the data are not already formatted and documented according to best practices). The CESM Chief Scientist and the CESM Data Manager should be contacted during proposal preparation to discuss possible (but rare) exceptions.

CESM Data Produced by University Projects

An increasing number of university PIs use CESM to perform simulations for their projects funded by various agencies. Even though the datasets from some of these simulations may be of benefit to the broader CESM community, the CESM Project does *not* have the storage resources to curate and maintain data generated from such projects. As stated above, the CESM Project's storage allocation at NCAR is limited to those simulations that are produced on the NCAR supercomputers by the CESM Project allocation.

Procedure

The CESM Project cannot entertain requests for storage of CESM datasets that result from agency-funded university projects, although the procedures in this document can be adopted as best practices for preparing CESM datasets for maximum value to the broader community. University PIs are expected to archive and make available any CESM data that are needed to reproduce their publications via other channels, such as campus repositories, pangaea.de, zenodo.org, and bco-dmo.org. However, NCAR is investigating the possibility of expanding its data curation and stewardship services to support the needs of a broader university community. Such a service would also include a system usage rate that would enable external data producers to pay for storage services on CISL infrastructure. University PIs are encouraged to communicate their interest in such a service to their NSF program manager and to NCAR (via alloc@ucar.edu).

Data Repositories at non-NCAR Facilities

The CESM Project may carry out simulations at a number of national and international computing centers, that is, at non-NCAR facilities, in partnership with its collaborators. These simulations, such as the CESM2 Large Ensemble, usually generate large volumes of data. The CESM Project's storage allocation at NCAR resources is limited to those simulations that are produced on the NCAR supercomputers. Hence, no additional storage allocation is available for simulations run elsewhere by the CESM Project. Therefore, it is necessary to discuss and clarify up front which data will be archived and how they will be distributed by the non-NCAR center, following the CESM data retention and distribution policies. This is particularly important for Public data. The CESM Project expects that the non-NCAR center has the storage and distribution means for these CESM datasets.

Procedure

CESM data generated at non-NCAR facilities should be archived at either the site of generation or its associated data archive center following the CESM procedures. Similarly, these data should be served to the broader community by the non-NCAR facility. The PIs should provide the filenames and metadata for such datasets to the CESM Data Manager to provide relevant information to the broader CESM community. Under rare circumstances, a limited subset of data from these simulations may be archived at and served from NCAR if prior arrangements are made with both the CESM Chief Scientist and CISL management.

Cloud-Hosted Data Repositories

In conjunction with the Science at Scale team at NCAR, subsets of large, high-use datasets determined by PIs based on previous / anticipated use, e.g., from the CESM2 Large Ensemble, are published on cloud platforms, such as Amazon Web Services. The netCDF files are typically converted to a cloud-native and cloud-optimized data format, called Zarr. These datasets allow users to directly subset the data, without the need to download it directly; but rather streaming the data as needed for analysis.

Procedure

The CESM Project will continue to take advantage of emerging opportunities to serve datasets on various cloud-based platforms.

CESM Data Used in Publications

Procedure

Publications that use public CESM data should utilize digital object identifiers (DOIs) to cite the output used in the publication. DOIs can be minted by the CESM Data Manager or the GDEX coordinator as needed.

The CESM Project is not responsible for serving non-Public CESM datasets used in publications to meet journals' data availability requirements. Such datasets may include data derived from Public CESM data and data generated by CESM simulations performed by other users in support of their funded projects.

Appendix A: CESM Data

CESM requires a set of initial and boundary condition files to start an integration. It then produces three distinct output data streams: plain-text log information, restart data, and history data. After an integration or case finishes, the raw history data are post-processed into more useful collections referred to as post-processed history data. A summary of these data files along with their descriptive size and format is presented in Table A1.

Table A1. CESM input and output data types, their descriptive size, and format.

Data Description	**Volume**	**Data Format**
a. Input Initial and Boundary Condition	small	netCDF, raw binary, and ascii
b. Output Log	small	Plain text files
c. Output Restart	medium	netCDF and raw binary
d. Output Raw History	large	netCDF
e. Post-processed History	large	netCDF, other (e.g., Zarr)

a. Input Initial and Boundary Condition Data

CESM runs are typically started using initial data that represent a known or idealized climate state for each component. Boundary condition files may also be used to prescribe (usually) time varying values of variables that are not predicted, such as the annual cycle of ozone in the atmosphere; land use and land cover change in the land; and emission datasets for future climate change scenarios. As these data are used repeatedly by users and their volume is relatively small, the data retention period for these data is indefinite.

b. Output Log Data

The output log data contain diagnostic messages written by each component during an integration, including a plain-text log file for the entire system. These log files serve to archive details about the model run, containing information about the length of the run; when it stopped and restarted. While the log output contains little information on detailed model diagnostics, it provides a convenient method for displaying quick-look diagnostics. Log data should be kept for a minimum of 4 years, after which the data can be deleted by the CESM Data Manager in consultation with the relevant PI(s).

c. Output Restart Data

While the majority of the CESM restart data files are in netCDF, there are also raw binary restart files. The restart data contains sufficient information for the model to restart exactly provided that the model code base or compilers have not changed. Restart data are usually output every two model years, although more frequent restarts are obtained with increasing model resolution. Retention periods of restart data should align with the procedure outlined in the Data Retention section for the post-processed history data, but exceptions may be required or considered by the CESM Data Manager in consultation with the relevant PI(s).

d. Output History Data

Raw CESM history data are the original data streams directly created by each component during an integration. These data consist of grid point representations of three-dimensional (latitude, longitude, time) and four-dimensional (latitude, longitude, height/depth, time) model fields. These fields include such variables as surface temperature, precipitation, and ocean salinity. Output frequencies can range from minutes to months or years, and the data can represent, for instance, instantaneous values, minimum and maximum values, or time-average values (daily-, monthly-, annual-means) over the output period. In total, hundreds of fields can be output by each component. Raw CESM history data should be deleted at the latest within 1 year of their creation, but preferably immediately after the creation of the post-processed timeseries format data, subject to the verification of the latter set.

e. Post-Processed Data

Post-Processed data are all other CESM data products. The component models are optimized for very high-speed multi-processor computing and output. This results in output streams from each component that may not be optimal for efficient retrieval and analysis. While raw history data can be used, time series analysis of such data presents challenges, necessitating access to the entire output volume of interest. The process of transforming the raw CESM history output into data collections more useful for analysis is called post-processing. This step may involve reformatting the data, deriving new fields from the existing data, making averages along any or all of the data dimensions, or sampling the data in different ways. These post-processed data are usually much smaller in volume and thus allow for more efficient retrieval and use. A major effort that produces the majority of the CESM post-processed files is the standardization of the raw datasets to comply with the requirements of the Climate Model Output Rewriter (CMOR) for submission to the CMIP. This process also requires renaming of many output fields. The retention periods of these post-processed data are described in the Data Retention section.

Appendix B: Data Tools

The CESM project uses primarily netCDF as its output data format and benefits from the large suite of software tools that support this format. Unidata (http://www.unidata.ucar.edu) has an extensive listing of software that can manipulate netCDF data. Currently, each model component maintains its own post-processing utility suite that can be accessed from the release code repository. Component post-processing utilities are provided only as a service to the community, so only informal community support is provided via the CESM bulletin board. See http://www.cesm.ucar.edu/models/cesm2/model_diagnostics/ for more information.

In addition, the Earth System Data Science (ESDS) initiative at NCAR is working to develop new Python-based tools to work with CESM data assets. These tools leverage the existing scientific Python stack, creating new tools when needed. For the latest updates regarding these tools and their availability, see the ESDS website at https://ncar.github.io/esds/.