BAMS
Meeting Summary

Road Map for the Next Decade of Earth System Reanalysis in the United States

Sergey Frolov, Cécile S. Rousseaux, Tom Auligne, Dick Dee, Ron Gelaro, Patrick Heimbach, Isla Simpson, and Laura Slivinski

What: Earth system scientists and Earth data users identified requirements, challenges, and opportunities for the next generation of the Earth system reanalysis products in the United States.
When: 16–18 May 2022
Where: Boulder, Colorado, and online

KEYWORDS: Atmosphere; Ocean; Carbon cycle; Reanalysis data; Air quality; Data assimilation

https://doi.org/10.1175/BAMS-D-23-0011.1
Corresponding author: Sergey Frolov, sergey.frolov@noaa.gov
In final form 31 January 2023

AFFILIATIONS: Frolov—National Oceanic and Atmospheric Administration, Boulder, Colorado; **Rousseaux and Gelaro**—National Aeronautics and Space Administration, Greenbelt, Maryland; **Auligne**—Joint Center for Satellite Data Assimilation, Boulder, Colorado; **Dee***—Planet A Consulting, Hudson, New York; **Heimbach**—The University of Texas at Austin, Austin, Texas; **Simpson**—National Center for Atmospheric Research, Boulder, Colorado; **Slivinski**—Cooperative Institute for Research in Environmental Sciences, Boulder, Colorado
*** CURRENT AFFILIATION:** Planet-A Consulting OÜ, Tallinn, Estonia

Reanalysis combines historical observations with modern Earth system models (ESMs) to generate a spatially and temporally complete history of the Earth system. It is an essential infrastructure that supports mission-critical activities across multiple U.S. agencies [including the National Oceanic and Atmospheric Administration (NOAA), National Aeronautics and Space Administration (NASA), National Science Foundation (NSF), Department of Energy (DOE), and Department of Defense (DOD)], industry (including energy, resource management, agriculture, infrastructure, insurance, information technology, and finance), and academia. In particular, reanalysis products can be used as initial conditions to evaluate and calibrate environmental forecasts produced by NOAA, NASA, and DoD and research institutions that investigate predictability on subseasonal to decadal time scales. Reanalyses can also provide an essential climate record of past weather conditions, extremes, and trends and can serve as verification datasets for ESMs. Reanalyses can quantify storage within and fluxes across the Earth system components essential to livelihood and commerce, such as heat, radiation, water, air quality, and carbon. Most recently, reanalysis datasets are also used for training of the machine learning models and as critical ingredient of emerging digital twins of the Earth system.

In May 2022, the U.S. Climate and Ocean: Variability, Predictability, and Change Program (U.S. CLIVAR) convened an international workshop to lay out a vision for the next decade of Earth system reanalysis efforts in the United States. The leading theme of the workshop was the evolution from reanalyses that include only a single component of the Earth system (such as the atmosphere or the ocean) to the production of multicomponent and possibly coupled Earth system reanalyses. Most importantly, this new generation is envisioned to be "consistent" (to be defined below) across multiple components of the Earth system, such as atmosphere, ocean, ice, land, atmospheric composition and air quality, carbon stocks, and the hydrological cycle. Ideally, the next-generation reanalyses would yield consistent fluxes between components and the ability to close essential budgets of properties that are exchanged between components. These reanalyses should ideally also be consistent in time, i.e., avoiding spurious trends in the time series related to changes in historical observational data coverage and quality. The next generation of reanalyses should also include consistent and trustworthy estimates of uncertainty. Finally, the reanalysis products will employ consistent data storage and access patterns [including analysis-ready data, cloud-optimized (ARCO) data formats] that will allow users to engage with datasets using emerging computing technologies in support of open science.

A consensus was reached that striving for a single reanalysis product that integrates all components of the Earth system and satisfies the diverse user needs is infeasible and would likely degrade the accuracy of individual Earth system components. Furthermore,

availability of historical data imposes different limits on the practical backward extension of each of the components. Instead, a hierarchical development approach was identified as a viable strategy to produce a suite of consistent reanalysis products (see Fig. 1). The backbone of the suite would include sparse-input centennial reanalysis (that only assimilates surface observations with a long historic record) and a full-input modern-era reanalysis (that assimilates all available data, including satellite records only available since the late 1970s), each produced with state-of-the-art coupled atmosphere, ocean, ice, and land models. These backbone products then can support production of downstream products such as carbon stock, air quality, hydrological, or biogeochemical reanalyses at global or regional scale.

To enable production of consistent datasets, a sustained, multiagency investment in a shared infrastructure will be necessary. This includes a common access pattern for products generated by multiple institutions, as well as user support and education. Such framework should be consistent with emerging computing infrastructure (e.g., cloud based) and available to the wide spectrum of scientific and commercial users. To support shared production and experimentation across institutions, a shared database of reanalysis inputs is needed, including historical observations and model boundary conditions such as land-use change and emissions. Greater efficiency can be further achieved by sharing component models (e.g., FV3 for the atmosphere, MOM6 for the ocean, CICE for polar sea ice, or MITgcm for 4DVar-enabled reanalysis) and data assimilation systems [e.g., the Joint Effort for Data assimilation Integration (JEDI) or open-source automatic differentiation tools for flexible adjoint code generation] required for reanalysis. Development and production of the reanalysis suite will require substantial investment in research and development of advanced coupled data assimilation algorithms, and in shared computing resources (see Fig. 1 for estimated core hours required).

To develop, produce, and effectively distribute reanalysis products that address future needs of the U.S. government, research, and commercial sectors, a sustained and coordinated effort across multiple U.S. agencies is needed. A related effort in Europe is conducted under the European Union's Copernicus Earth Observation Programme, which provides long-term funding for the development of reanalysis products as an operational service. The annual investment from the European Commission in this effort is considerable, supporting the development, production and dissemination of global and regional reanalysis datasets,

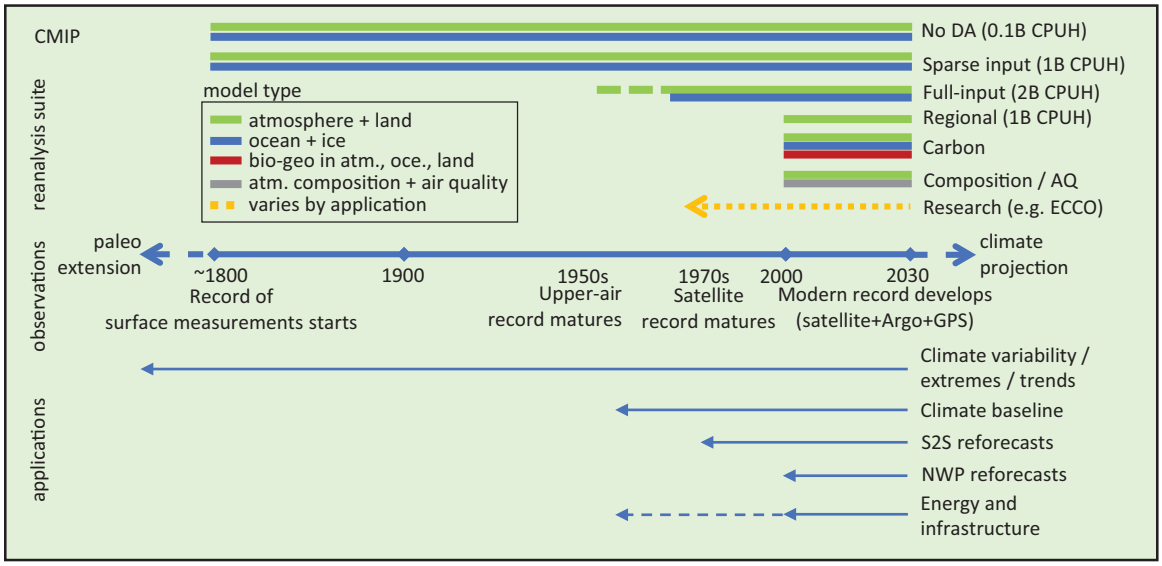

Fig. 1. Proposed suite of reanalyses and its relationship to available observational data and application drivers. 1B CPUH stands for an equivalent of 1 billion hours of computation on a single computational processing unit.

as well as various activities related to training, user support, and user engagement. It also supports essential work on improving the availability and quality of input observations for reanalysis, such as reprocessing and quality control of existing satellite data records as well as data rescue and harmonization of in situ data sources from observing networks and the archives.

In the United States, such investments are distributed across multiple agencies. For example, NOAA has a mandate to maintain a record of past climate and serve as an authoritative source of information about future weather and climate. NASA is invested in placing its vast suite of Earth observations in a climate context, with dedicated foci on tracking global and regional sea level change, as well as developing a global carbon monitoring system. DOE is the leading provider of the supercomputing resources in the United States and develops novel computational tools for advanced Earth system models. The U.S. Army Corps of Engineers has a mandate to effectively manage over 50% of water storage across the United States. And U.S. Environmental Protection Agency regulates and studies air quality. No single U.S. agency possesses the expertise, the resources, or the critical mass to enable the full production of the reanalysis suite depicted in Fig. 1. *Successful fulfillment of the vision for the consistent reanalysis outlined in this document represents a grand computational, organizational, and scientific challenge for the next decade.* To achieve the vision outlined in this report, a mechanism for interagency collaboration and resource sharing will be needed.

Workshop objectives and format

The workshop brought together experts from data assimilation communities, operational and research centers, academia, and (research and private sector) the reanalysis user community. Specific goals of the workshop included the following:

- Identify scientific goals and requirements for the next generation of integrated Earth system reanalysis that includes the atmospheric, oceanographic, and cryosphere perspectives, as well as the cycling of important tracers between these components.
- Identify priorities and opportunities for tighter collaboration between U.S. and international reanalysis communities.
- Review existing infrastructures for generation, storage, and distribution of reanalyses in the United States.
- Develop strategies to account for and minimize biases, spurious trends, and property nonconservation in reanalysis.
- Review algorithmic requirements and innovations to improve integrated or coupled data assimilation, and to produce trustworthy uncertainty estimates.
- Explore the use of reanalysis approaches for comprehensive model calibration.
- Develop strategies for tighter coordination between observation and reanalysis communities.
- Identify priorities and develop a research road map for coupled reanalysis.
- Evaluate strategies for including aerosols, ocean and terrestrial biogeochemistry, and atmospheric composition as a fully coupled part of the reanalysis.
- Establish a tighter link between the reanalysis, prediction, and projection enterprises (e.g., seasonal–decadal–CMIP).

Discussions were conducted throughout the 2.5-day meeting using a combination of plenary and small-group breakout sessions. The first day focused on the scientific and application-based requirements for consistent climate reanalysis. The second day targeted technological opportunities and limitations. Finally, the third day focused on opportunities

for collaboration. A series of webinars were conducted in the 4 months leading up to the workshop.

The slide decks and the recordings from the webinars and the in-person meeting are available at https://usclivar.org/meetings/reanalysis-2021.

Workshop findings

Session 1: Scientific and application-based requirements for consistent climate reanalysis. During the first day of the workshop, invited speakers presented a variety of perspectives on the scientific and application-based requirements for the next generation of reanalysis products. The need for "consistent climate reanalyses" emerged and was present throughout the breakout discussions. We define the concept of consistency in this section.

Depending on the users, the requirements for a "consistent climate reanalysis" varied greatly. For example, those who wish to examine long-term trends are looking for consistency in analysis procedures and input datasets over time, while those who use reanalysis products for initialization of reforecasts are looking for the most realistic state possible. The discussion on this topic made clear that it would be a challenge to produce a single consistent reanalysis that will satisfy all user needs. Instead, a suite or hierarchy of reanalyses may be desirable.

For the use of reanalyses for investigating climate anomalies (e.g., heatwaves) or long-term climate trends, there is a need to minimize artifacts due to observing system changes, artifacts in the DA scheme, and spinup at the beginning of different streams. This can be defined as *consistency in time*. Several techniques based on machine learning were presented during the workshop that have a potential to minimize such artifacts. There is also a clear desire for reanalyses to estimate realistic anomalies and trends for the right reasons in order to use these products to gain a process-level understanding (defined as *dynamical consistency*). Increased utilization of budget terms in the conservation equations for tracers and the role of analysis increments may help to address this problem through improved understanding of the processes involved and model biases in the underlying forecast system. There may be much to gain from increasing the access and usability of data products intended for examining budgets and analysis increments across reanalysis centers, e.g., by using *consistent data formats* and making a greater suite of fields available.

There is also a need for improved *consistency across modeling components*. This is particularly true when considering the full Earth system. For example, the CO_2 field that is input to the radiation scheme of the atmosphere is typically inconsistent with the fields determined using the carbon modeling components of the models. Another important, and related, aspect that would benefit greatly from increased consistency is the representation of carbon fluxes between the ocean, land, and atmosphere; this is particularly important for carbon accounting. A third example addresses the energy transfer between the coupled atmosphere–sea ice–ocean system. There is likely much to be gained by increased collaboration with Earth system model development communities in this regard.

Findings for session 2: Technological opportunities and limitations. To determine the best path forward for achieving the 10-yr goal of a consistent Earth system reanalysis, limitations and opportunities within each of the three necessary components of a reanalysis (observations, model, and data assimilation method) were identified. To improve future reanalyses, historical observations must be acquired through a process known as "data rescue," and existing observations may need to be continuously reprocessed (particularly for satellite observations). Participants discussed the need for an organized effort toward data rescue and reprocessing, especially for historical satellite observations, and increased effort to create a centralized database of observations of atmosphere, ocean, land, and air quality.

Participants also discussed opportunities for collaboration between the reanalysis and model development communities, since reanalysis development can both inform and benefit from model development. Examples of possible collaboration include machine learning–based bias correction algorithms applied during reanalysis production as well as diagnostics applied to the final reanalysis product. While these examples stem from the numerical weather prediction (NWP) community, reanalysis efforts should be separated from NWP, since reanalysis should itself be considered an essential climate service.

Different data assimilation methods for coupling across components were presented, including weakly coupled DA, strongly coupled DA, and coupled replay. In this context, there are several reanalysis development strategies. While many agree that a single reanalysis that is coupled across all Earth system components would be ideal, it is currently infeasible to achieve without significant degradation to the reanalysis' performance. Alternatively, the United States should develop a hierarchy of reanalyses as illustrated in Fig. 1. Specifically, it was suggested that the United States should develop a set of reanalyses with systematically reduced observing networks, including full input (with satellite data), conventional only (no satellites), and surface only (no upper-air data), as well as a free-running model simulation without assimilating any observations This hierarchy could also include a suite of reanalyses that focus on different components of the Earth system, such as the carbon cycle coupled across the atmosphere, ocean, and land components. Another avenue is the development of coupled reanalyses that are property conserving, such that the interpretation of local and global tendencies, as well as fluxes across the components become physically unambiguous.

Quantifying uncertainty is a necessary feature of reanalysis development. Methods for quantifying uncertainty with ensemble DA algorithms as well as those exploiting derivative (Hessian) information were presented, and discussions focused on the need for a systematic approach to quantifying uncertainty to help both development of reanalyses as well as evaluation of reanalyses. Overall, consistent, (strongly) coupled Earth system reanalysis with quantified uncertainties offers a range of prospects for scientific and commercial applications, but it remains a grand challenge algorithmically and computationally.

Consistent, (strongly) coupled Earth system reanalysis with quantified uncertainties offers a range of prospects for scientific and commercial applications, but it remains a grand challenge algorithmically and computationally. The challenges are closely connected to challenges in developing digital twins of the Earth system. The digital twin (DT) concept has garnered much momentum in recent years, with examples being the Destination Earth[1] (DestinE) project supported by the European Commission, the U.S. National Academies' Committee on the Foundational Research Gaps and Future Directions for Digital Twins,[2] the WCRP Digital Earth Lighthouse Activity,[3] the UN Ocean Decade Digital Twin of the Ocean (DITTO) program,[4] or private-sector efforts such as NVIDIA's Earth-2.[5] In particular, three characteristics of the digital twin concept are synonymous to those of Earth system reanalysis workflows discussed here: (i) the DT must improve through the integration of new data and provides a dynamic digital history of the asset or entity with quantified uncertainties; (ii) the DT must entail synergistic two-way coupling between the physical system, the data collection, and the user/social system; and (iii) DTs will harness emerging high-performing computing hardware architectures. Arguably, data assimilation workflows developed for Earth system prediction and reanalysis, potentially enhanced with emerging approaches from scientific machine learning and reduced-order modeling will be at the heart of successful, global-scale digital twin efforts.

[1] https://digital-strategy.ec.europa.eu/en/policies/destination-earth

[2] www.nationalacademies.org/our-work/foundational-research-gaps-and-future-directions-for-digital-twins

[3] www.wcrp-climate.org/digital-earths

[4] https://ditto-oceandecade.org

[5] https://blogs.nvidia.com/blog/2021/11/12/earth-2-supercomputer/

During the breakout session discussions, participants identified several strategies that could reduce the cost of the reanalysis development and production. These include

- a hierarchy (or suite) of reanalyses as a strategy for achieving consistent reanalyses, helping address computational cost;
- more organized collaboration (such as a working group) to address data rescue needs, particularly for satellite data;
- common infrastructure for reanalysis inputs (including forcing observations and boundary conditions) and outputs (including analysis fields, tendency terms, analysis increments, observation diagnostics, and analyzed climate indices), possibly exploiting cloud resources and corresponding cloud-optimized data formats;
- agreement on common metrics and diagnostics to aid comparisons across reanalyses, possibly coordinated with diagnostics employed by the CMIP community; and
- agreement upon a systematic approach for quantifying uncertainty.

Findings for session 3: Collaborations. Summarizing the findings of the first 2 days of the workshop made it clear that no single U.S. agency has sufficient mandate, knowledge, human, or computational resources to fulfill the vision of the reanalysis suite presented in Fig. 1. As a result, a mechanism for interagency collaboration and coordination is needed. One fruitful approach within the U.S. community would be to focus on a shared input database. Such an input database would include observations, model forcing (e.g., emission and climate forcing), and boundary conditions (e.g., SST, sea ice, land-use, and river discharge databases for partially uncoupled reanalyses). In addition, new investments in common formats and for sharing of the reanalysis output and diagnostics are needed. This will be a prerequisite to coordinated experimentation to explore the trade space of configurations for coupled products. Such coordinated experimentation should include

- continuous refinement of the observational archives through reprocessing and rescue of historic observations;
- development, testing, and sharing of observational operators and quality control procedures for historic observations;
- characterizing and reducing uncertainty in long-term trend estimates by accounting for artificial jumps in the mean state due to changes in the configuration of the observing system; and
- developing cost-effective strategies to produce a consistent reanalysis across multiple components of the Earth system.

Another shared investment that is well underway across a wide swath of the U.S. modeling and reanalysis community is the investment in shared modeling and data assimilation components, such as the FV3 atmospheric model, MOM6 ocean model, CICE ice model, GOCART atmospheric composition model, 4DVar-based automatic-differentiation enabled MITgcm, JEDI data assimilation, and ESMF model coupling frameworks. However, as more advanced reanalysis products become available (such as carbon stock estimates or hydrological reanalysis) a broader engagement with the wider user community is needed to test and diagnose these new products. This can include shared metrics for evaluating quality and consistency of the reanalysis products and a common store and application programming interface (API) for accessing reanalysis products across different generations of products and agencies. While Copernicus Climate Service provides an inspiration for such a single data store, a future U.S. implementation should engage more fully with cloud providers that can help collocate the data store, the APIs, and the computing resources for the end users of the reanalysis products.

A "scientific data commons" could leverage both commercial cloud computing and HPC centers for analysis of the same shared data. Very high bandwidth connectivity at the national level between data commons, cloud regions, and HPC centers would unleash Earth system data access and analysis capabilities.

Recommendations

Recommendations from the reanalysis workshop can be grouped into three categories: 1) definition of a consistent Earth system reanalysis and the hierarchical development approach presented in Fig. 1, 2) specific scientific challenges for the next generation of reanalyses, and 3) needed investments in the shared infrastructure for collaboration.

Attributes of consistency that should define the next generation of the Earth system reanalysis suite include

- consistent in time and across observing systems,
- consistent error estimates,
- consistent fluxes and budgets,
- consistent API for accessing the data, and
- consistent error metrics to track and guide improvements in reanalysis.

Specific scientific challenges that should be addressed for the next generation of reanalysis include the following:

- The next generation of the backbone sparse and full-input reanalysis will rely on fully coupled models of the atmosphere, ocean, land, and ice. Representation of consistent budgets (balance of storage, transports, and fluxes) and reduction of coupled model biases is a well-recognized challenge in the community.
- Accounting for storage and fluxes of heat, water, and carbon is essential for climate monitoring services. Representation of the carbon stocks and fluxes in the land components needs focused attention and coordination across the U.S. research community.
- Representation of precipitation, droughts, extreme events, and water movement between atmosphere, land, cryosphere, and ocean presents a fundamental challenge for the current generation of coupled models.
- A more realistic representation of tropospheric ozone will be needed to reconstruct historical air quality. This will require improved databases of emissions that are consistent across land, atmosphere, biosphere, and anthroposphere.
- Coupled processes across the atmosphere–ice–ocean system in polar regions are particularly challenging to simulate and poorly constrained by observations, requiring focused efforts for improving reanalyses.
- Finally, to reduce the impact of coupled biases and imbalances between model components, the modeling and data assimilation communities need to incorporate knowledge of systematic model errors gained from the assimilation in the model development process. Furthermore, data assimilation frameworks should be extended to tackle comprehensive model parameter calibration, which holds promise to contribute to model error/bias reduction.

To enable effective progress toward the next generation of reanalysis products, a sustained investment into joint infrastructure for collaboration is needed. Aspects of this investment include

1) collaboration on sharing of reanalysis inputs, including shared observations, land-use change, and databases of emissions and fires;
2) common infrastructure for sharing reanalysis output, including standardized API for gridded fields, analysis-ready cloud-ready data formats, observational feedback, bias estimates, tendency terms for tracers, and derived metrics;
3) common data access patterns that will enable collocation of end-user computations with the data products from a variety of U.S. and international agencies, e.g., via a scientific data commons paradigm;
4) collaboration on data rescue and development of observational operators, bias correction, quality control procedures, and observation reprocessing;
5) shared resources for support and education of end users about most appropriate use of the reanalysis resources;
6) infrastructure for shared experimentation, including continuous investment in common computational algorithms and software components for models, data assimilation, and data access systems;
7) exploration of machine learning approaches at various stages of the reanalysis production workflow;
8) sufficient computational resources to enable development of reanalysis products across the spectrum of the U.S. agencies;
9) a forum and funding mechanisms to encourage interagency collaborations; and
10) integration of the reanalysis enterprise as a key ingredient of the emerging digital twins of the Earth system.

Acknowledgments. U.S. CLIVAR provided financial and organizational support for this meeting.