Earth System Reanalysis in Support of Climate Model Improvements

Detlef Stammer,[a] Daniel E. Amrhein,[b] Magdalena Alonso Balmaseda,[c]
Laurent Bertino,[d] Massimo Bonavita,[c] Carlo Buontempo,[c] Nico Caltabiano,[e]
Francois Counillon,[d] Ian Fenty,[f] Raffaele Ferrari,[g] Yosuke Fujii,[h] Shreyas Sunil Gaikwad,[i]
Pierre Gentine,[j] Andrew Gettelman,[k] Ganesh Gopalakrishnan,[l] Patrick Heimbach,[i]
Hans Hersbach,[c] Chris Hill,[g] Shinya Kobayashi,[m] Armin Köhl,[a] Paul J. Kushner,[n]
Matthew Mazloff,[l] Hisashi Nakamura,[o] Stephen G. Penny,[p] Laura Slivinski,[q]
Susann Tegtmeier,[r] and Laure Zanna[s]

KEYWORDS:	**Improving Climate Models and Projections Using Observations Workshop 2023**
Data assimilation; Model errors; Model evaluation/ performance; Model initialization; Numerical weather prediction/ forecasting	*What*: More than 70 experts from research and operational centers came together in a hybrid meeting held at the Massachusetts Institute of Technology (MIT) to discuss the state of climate-model-related data assimilation and to inspire future work required to make coupled data assimilation and Earth system reanalysis a reality.
	When: 12–14 June 2023
	Where: Massachusetts Institute of Technology, Cambridge, Massachusetts

AFFILIATIONS: [a] Universität Hamburg, Hamburg, Germany; [b] NSF National Center for Atmospheric Research, Boulder, Colorado; [c] European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom; [d] Nansen Environmental and Remote Sensing Center, Bergen, Norway; [e] World Meteorological Organization, Geneva, Switzerland; [f] NASA Jet Propulsion Laboratory/Caltech Pasadena, Pasadena, California; [g] EAPS, Massachusetts Institute of Technology, Cambridge, Massachusetts; [h] Meteorological Research Institute, Japan Meteorological Agency, Tokyo, Japan; [i] Oden Institute, The University of Texas at Austin, Austin, Texas; [j] Columbia University, New York, New York; [k] Pacific Northwest National Laboratory, Richland, Washington; [l] University of California, San Diego, California; [m] Japan Meteorological Agency, Tsukuba, Japan; [n] University of Toronto, Toronto, California; [o] The University of Tokyo, Tokyo, Japan; [p] Sofar Ocean, San Francisco, California; [q] NOAA/OAR/Physical Sciences Laboratory, Boulder, Colorado; [r] University of Saskatchewan, Saskatoon, California; [s] New York University, New York, New York

1. Assimilation workshop summary

A 3-day workshop took place from 12 to 14 June 2023, at the Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, focusing on data assimilation (DA) and machine learning (ML) in the context of Earth system reanalysis and climate model improvements.

The workshop, organized 25 years after the inception of the Estimating the Circulation and Climate of the Ocean (ECCO), was an effort to lay out the roadmap for future development of DA in support of climate modeling and climate knowledge improvements, or "climate DA." The following is a summary of the workshop outcomes and recommendations arising to move the field of DA forward in the context of climate modeling.

Recent climate model developments, established through increased model resolution, have led to substantial improvements in model simulations of the time-evolving, coupled Earth system and its subcomponents. However, regardless of resolution, climate models will always produce climate features and variability that differ from the real world and will be prone to biases. This is due to many remaining uncertainties, such as in parametric and structural model uncertainty, in the initial conditions prescribed, and in the prescribed (scenario) forcing which varies on decadal to centennial time scales.

Further model improvements are expected to arise specifically from the improved representation of physical processes realized through model–data fusion. This will create an unprecedented opportunity to better exploit a large array of Earth observations, from in situ measurements to weather radars and satellite observations, as the resolved scales of the models approach those of the observations. For this, climate DA will be the central tool to bring models and observations into consistency, by improving initial conditions, inferring uncertain model parameters and structure, and quantifying uncertainty. Generally, there will be advantages and complementarities of adjoint-based smoother approaches, ensemble-based filter approaches, or new ML-inspired approaches. Yet the ever-increasing model resolution will present growing challenges arising from computational cost, calling for new ways of performing data assimilation and model optimization. Using the complementarity in a hybrid approach, blending tools and concepts from variational, ensemble, and ML methods might be what is required in the future. In this context, ML could be important to handle nonlinear responses and to better approximate non-Gaussian distributions.

The central goals of climate DA are to provide (i) the best possible description of the time-evolving state of the climate over the instrumental era, thereby optimizing the use

and interpretation of climate observations, (ii) well-balanced initial conditions that may serve as suitable initial conditions free of initialization shocks for skillful prediction, and (iii) observation-constrained estimates of model parameters and biases. Climate DA must aim to enhance climate knowledge through the improved ability to simulate and predict the real world by optimally combining Earth system models and the most available global observations from different Earth system components and domains. For this, climate DA needs to bring the simulations of climate models into consistency with the natural world as observed by the global climate observing system and to produce a dynamically balanced climate estimate in support of initialized climate predictions, investigation of climate processes, and the identification and reduction of model bias. In the future, arguably the most important aspects of climate DA and ML in support of model improvement and enhancing predictive capabilities might become optimizing initial conditions, model parameters, and model structure to mitigate model biases and thereby improve models' skill in simulating the observed climate, as well as to enhance model skill for climate projections.

Multiple atmospheric reanalyses and the more recent emergence of ocean or coupled ocean/sea ice reanalyses are prominent examples of such efforts. However, although many Earth system component-focused reanalyses exist and have enabled enormous progress, there remain significant gaps in progress toward fully coupled Earth system reanalysis. This gap is exemplified by the disconnect between ongoing reanalysis efforts, for the atmosphere typically conducted at operational prediction centers, and the climate modeling enterprise as conducted under the auspices of the World Climate Research Programme (WCRP) Earth System Modeling and Observations (ESMO) core project. A major thrust of the workshop was therefore to point to the need to close this gap through new efforts targeting full Earth system reanalyses. Translating lessons learned from ultra-high-resolution models to lower-resolution models used for lower-resolution climate projections and vice versa is yet another area where exploration in climate modeling and DA is needed. These improvements eventually will likely have benefits also to the forecasting and numerical weather prediction communities.

2. Workshop recommendations and future plans

Enhancing and further developing DA methodologies, including parameter estimation approaches, ML tools, and model structure inference, will be essential to further exploit the availability of new climate modeling tools and high-resolution observations and to provide a detailed analysis of the state of Earth. The field must now target a comprehensive approach toward coupled Earth system reanalysis and better future projections by improving the description of changes of energy, water, carbon, and biogeochemical cycles in the system and by allowing for more consistent analyses of climate variability and change, taking into account feedbacks and interactions. Accomplishing physical and dynamical consistency in the assimilation and estimation process is key for reaching at least some of these goals. Work must also be extended to include biogeochemical cycles.

To realize the full capabilities of climate DA, we need to advance science and technologies for analyzing and merging global observations and Earth system model data in the context of Earth system DA. A close collaboration of DA scientists with model developers is needed to enable parameter and model structure improvements and model bias correction by applying this new technology. Such an effort would be facilitated by a tighter connection between the assimilation community and Earth system modeling that is beyond the capacity of current infrastructures. Automatic differentiation (AD) can be the backbone for the integration of modern climate model code and its optimization. This includes the development of future models in new software paradigms or programming languages that support AD. It is also important

to highlight the need for maintenance and expansion of the global climate observing system, as well as incorporating new components, via instrumental and algorithmic technological advances. This requires a proper quantification of uncertainties arising from changes in the observational records and from other error sources.

An optimal fusion of models and observations will also enable us to determine where observations are most needed to reduce uncertainty and enable the enhancement of the observation system. This can be accomplished through Observation System Simulation Experiments (OSSEs) or other quantitative observing network design approaches. This will also improve understanding of climate variability and feedback mechanisms as well as predictability on seasonal to interannual to decadal scales and beyond. Embedding this approach in an ensemble setting accounts for natural variability arising from chaotic dynamics and provides useful uncertainty estimates for the interpretation of Earth system observations.

Specific recommendations for required actions and funding that emerged from the workshop's discussions will be summarized below under the headings of 1) Exploitation of Earth Observations; 2) Development of AD and ML Infrastructure; 3) Advances in Ensemble and Variational DA and ML Methods; 4) Model Improvements through Parameter Estimation; 5) Performing Earth System Reanalysis; and 6) Understanding, Prediction, and Projection of Earth System Evolution.

a. Exploitation of Earth observations. Data assimilation fundamentally depends on the availability of calibrated and quality-controlled observations. However, much more information is required from observations prior to performing DA. At the same time, assimilation efforts can feed back valuable information about the observing system and its required modifications. In the context of climate DA, several aspects require attention:

- Observational data streams needed for DA: Building on the existing Observations for Model Intercomparison Projects (Obs4MIPs) (https://pcmdi.github.io/obs4MIPs/), observational data streams need to be developed and maintained for DA and reanalysis activities: Move Obs4MIPs to Obs4Assim to provide observations at various processing stages (L1, L2, L3, …), including those that are as close as possible to the raw data (e.g., profile data in the ocean, or along-track data from satellites) plus uncertainties, as well as software for transforming raw observations to common representations (e.g., well-documented inversion models/operators of satellite radiances to temperature that can be used within DA systems). This recommendation applies to all Earth system components and associated observational systems.
- Reference datasets for assimilation efforts: Reference datasets should be established that can be used by several groups as input to their assimilation efforts. For these datasets, existing inconsistencies should be removed. Datasets for cross-component (such as air–ocean, ice–ocean, land–ocean, or seafloor–ocean) fluxes should be included. Such reference datasets also need error information (see "correlation matrices" and "error information" below).
- Exploiting untapped variables in the suite of available observations: While several parameters are widely used in the study of Earth system (e.g., temperature and salinity), many variables have been observed and archived that are not yet used extensively to study Earth. The full spectrum of information available in data archives needs to be exploited (e.g., cloud or aerosol information) and the use for Earth system modeling and climate data assimilation expanded.
- Data archeology and data rescue: There are many past observations that have not been tapped into, partly because they remain to be digitized, quality controlled, and/or

calibrated. No other way will enhance the data density of the observations from the vastly undersampled past than data archeology and data rescue, which requires more attention and financial support.

- Correlation matrices: Often correlation and cross-correlation information of processes are required, e.g., in data studies, in modeling, or in assimilation approaches. Existing data archives need to be screened to infer and expand our knowledge about correlation and cross-correlation information.
- Error information (error covariances): A common understanding needs to be developed for errors and error covariances of existing or future observational products. This also includes cross-component fluxes available from reanalyses and should characterize systematic and random errors.
- Model–observation differences: The information content of space–time-dependent errors versus processes (e.g., ensemble spread) available from all model–observation comparisons and DA needs to be investigated with respect to model and/or observational errors, and respective information should be used to improve both, where possible.
- Optimizing climate observing systems: While a rich spectrum of Earth system observations exists today with unprecedented spatial coverage, such an observing system needs to be maintained and further extended (e.g., deep ocean observations, hydrological field station networks, and flux networks). With respect to climate observations, observational streams require dedicated cross calibration and stability, e.g., using suborbital missions to bridge gaps between remote sensing missions. Coverage of observations still needs to be optimized in space and time, e.g., to cover the deep ocean or the Arctic.
- Observation system evaluation and design: Further encouragement and funding is required for Observing System Experiments (OSE), OSSEs, and other observing system evaluation methods to predict and quantify observation impacts. Those activities provide scientific justification for maintaining and further expanding climate observing systems. They will also help to show where new observations would be most valuable for reducing uncertainty in climate projections. This would include advising on new missions and observables through interaction with the Global Climate Observing System (GCOS), the Global Ocean Observing System (GOOS), and the Committee on Earth Observations Satellites (CEOS).
- Interaction with GCOS, CEOS, and other data providers: While exploiting the information content of existing datasets and data streams, emphasis should be put on identifying important gaps in observing systems that need to be closed to improve DA results (e.g., more Argo observations in the Arctic and deep ocean, and more historical salinity data).

b. Development of AD and ML infrastructure. Performing Earth system–related DA requires the existence and availability of modern infrastructure and effective training. Beyond the existence of Earth system models, there is a need for open-source AD tools for the generation of adjoint codes and the availability of ML environments tailored to Earth system modeling applications that seamlessly integrate with climate model codes. The existing Earth system model environment needs to be able to deal with both. The following sustained developments are required:

- Open-source AD toolboxes: To expand the usage of DA activities using adjoint models, open-source AD toolboxes need to be developed, optimized, and maintained, capable of handling modern code differentiation and optimization.
- Open-source ML environments for Earth system simulation: ML tools need to be developed and maintained, capable of handling unstructured computational grids and sparse, indirect, and noisy observations.

- Hybrid DA approaches: Hybrid physics–based simulation, ML, and DA approaches need to be developed and made suitable for climate model calibration, parameter estimation, model structural bias reduction, and surrogate modeling.
- Cloud services in DA: The exploitation of cloud services needs to be considered to expand the available computational resources by developing systems that can use such cloud resources (public or private). Respective approaches need to ensure that accessing cloud data remains free or at a nominal cost.
- Software generation frameworks: The development of next-generation models in software frameworks (e.g., JAX or Julia) that natively support automatic differentiation toolboxes will be required. Such efforts should also harness evolving high-performance computing (HPC) architectures, such as an optimal use of GPUs for the purpose of DA.
- Capacity building: To help next-generation DA experts gain experience and to expand the community, guiding tutorials should be developed, such as "cookbooks" that also discuss good practices, that share information, etc. Educational programs should develop shared curricula to support teaching in this specialized field, such as by developing courses on automatic differentiation and hybrid modeling.

c. Advances in ensemble and variational DA and ML methods. In the context of coupled climate modeling and Earth system reanalyses, significant innovation in DA approaches and theoretical advances are required. This also applies to uncertainty assessment and more generally ways modeling can be driven or at least guided by observations:

- Uncertainty quantification: Develop or expand on theories and algorithms for quantification of uncertainty in any assimilation products, including those generated by adjoint models, ensemble Kalman inversions, and for ML algorithms.
- Multiscale approaches: Develop multiscale/multidomain coupled DA approaches required to address the large range of spatiotemporal scales present in coupled Earth system models. This effort needs to also consider the generation of error estimates.
- Information flow analysis: Improved understanding of the flow of information within assimilation methods is required to better understand solution trajectories and how they are constrained by observations. Such work could use adjoint sensitivities and would support the understanding of the dynamics and feedbacks involved in Earth system evolution.
- Advance in DA approaches: DA must be advanced and expanded. As part of this, we need to be able to
  - improve theory and methods to deal with sparsely observed nonlinear systems. An example is the ocean, which over long periods was vastly undersampled and still remains so below 2000-m depth;
  - deal with chaos and nonlinearities in the DA (e.g., synchronization and supermodeling) and its probabilistic nature, such as using generative ML algorithms;
  - merge ML (backpropagation) and traditional DA approaches, exploring the potential of hybrid assimilation methods (mix of smoothers, filters, and ML);
  - improve algorithmic developments of DA inspired by ML (backpropagation), leveraging AD and its capacity to handle nonlinear functions;
  - explore the use and limits of ML emulators within the DA cycle.
- Advance data-driven climate modeling: Further advance data-driven climate modeling and model improvements and make this a part of development workflows at major modeling centers. As part of this, develop and implement methods focusing on model parameter inferences and improvements.

d. Model improvements through parameter estimation. To reduce model biases, model physics needs to be improved by improving model structure and optimizing model parameters based on observations. In the context of advancing data-driven climate model improvements, it is especially the latter aspect that can lead to progress, although the former aspect can be addressed as well. Several steps are needed:

- Model structural improvements: Model structural improvements must be advanced; such an approach has to be model resolution dependent and should maximize the usage of information. Adjoint models are critical for this step. ML approaches used to replace model subcomponents being used in this way have shown great promise.
- Parameter estimation cannot just be done locally in space and time; it should make use of dynamical understanding and the propagation of information. This means that the tuning of parameters might occur across spatial regions and deal with the issue of dimensionality. Adjoint models are ideal to deal with this aspect by carrying information backward in space and time.
- Estimating new mathematical formulations: The estimation of unknown mathematical formulations must be advanced, e.g., through data-driven equation discovery (ML). The effort would also address improving formulations and parameters in (nonlinear) atmospheric or oceanographic boundary layers. Approaches should include missing components, such as wave models and fine-scale bottom topography.

e. Performing Earth system reanalysis. Performing a full Earth system reanalysis to include not only atmosphere/ocean/land/ice but also biogeochemistry (BGC) and carbon is unprecedented and entails many challenges that require coordinated and systematic attention. This next frontier in climate DA will require vastly more work than "cobbling together" existing components and workflows. It will need to involve the following aspects:

- Using adjoint modeling, the approach will require identifying sensible controls, e.g., bulk formulae, mixing parameters, cloud parameterizations, albedo, or aerosol feedback, and adjusting those parameters such that the solution better represents the observed state in some statistical sense.
- More control parameters and observations must be included which can constrain them [e.g., clouds, aerosols, soil moisture, land use and land surface coverage, bottom topography, ice shelves, loading and (self)attraction, and biogeochemistry].
- Processes that are currently often missing or misrepresented must be included, e.g., ocean and land biogeochemistry, wave coupling, carbon sources and sinks, and soil processes. This will require more interaction with the relevant communities.
- Investigation of new DA and ML methods in the context of Earth system reanalysis: Significant developments in reanalysis will require significant developments in coupled DA and ML that may be specific to reanalysis, not just numerical weather prediction. Hybrid approaches should also be investigated.
- Interaction with the in situ and satellite Earth observing community (GCOS, GOOS, and CEOS) is required to give feedback on the observing system and on the quality of existing data and areas where observations are missing.

f. Understanding, prediction, and projection of Earth system evolution. Earth system reanalysis data should be used to improve our understanding of processes and mechanisms, including coupled component interactions and feedbacks (e.g., terrestrial and marine boundary layer processes and coupled atmosphere–ice–ocean processes) and representation of climate forces. The experience of producing an Earth system reanalysis

should also inform how to improve models for forecasts at leads of subseasonal to seasonal (S2S) and seasonal to decadal (S2D):

- Clarifying the implications of the lack of direct observations of coupled processes, including interactions of physics and BGC. For instance, will a lack of observations limit the accuracy of estimating coupled processes?
- Better characterization of BGC, water and energy cycles, carbon cycle, budgets, and sources and sinks.
- Improving estimates of trends, extremes, and impacts. This applies to all modeling efforts and goes beyond assimilation.
- Coordinated updates to datasets of climate forcers to consistently drive Earth system reanalysis schemes. These forcings span natural and anthropogenic emissions, emissions related to short-lived and long-lived species, and boundary conditions from evolving land-cover and land-use changes.
- Estimating and understanding feedbacks and sensitivities to constrain climate sensitivity and transient climate response.
- Exploiting these improved dynamical prediction systems to produce climate predictions with credible uncertainties.
- Using systems to understand where additional observations can reduce uncertainty in climate predictions and projections

3. Concluding remarks

Improving our understanding and our capabilities to predict Earth system calls for a truly integrated Earth system modeling, calibration, analysis, and estimation framework. Such a framework must take advantage of the vastly expanding, yet diverse and heterogeneous observational data streams that measure the individual components. It must also go beyond a simple coupling of existing estimation systems by building a truly integrated system and involving multimodal approaches if it is to extract information contained in observations to fully elucidate coupled processes. The workshop made it clear that establishing such an Earth system modeling and reanalysis system represents a frontier in climate science involving modeling, observation, DA, and computational science communities. Achieving this frontier requires many steps, involving theoretical aspects of data assimilation, improving assimilation infrastructure, and improving data information extraction. Such a system can then also be used to continuously improve Earth system models while advancing knowledge about processes involved in the changes in the state and transport of heat, water, carbon, and biogeochemical constituents.

The use of observations for improving reanalysis, initialization, and models was a key theme of the workshop, and this is reflected in the recommendations provided above. Beyond these aspects, improved estimates of Earth system state (including but extending beyond initial conditions) will yield better climate information, namely, reanalysis, initialized predictions, and simulations of historical and projected climate. This also applies to the treatment of the temporal variations of boundary conditions, such as land cover, land use, vegetation, and aerosol emissions. Convergence on best practices will ultimately improve the quality, credibility, and therefore usability of climate information produced by the scientific community.