https://doi.org/10.5194/egusphere-2026-6437
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

The calibration deficit: a challenge for Earth System simulation and emulation

Peter M Cox[1], Dale Durran[2], Veronika Eyring[3,4], Tapio Schneider[5], Duncan Watson-Parris[6], Daniel Williamson[1], and Venkatramani Balaji[7]

[1]Department of Mathematics & Statistics, University of Exeter, Exeter, UK
[2]Atmospheric and Climate Science, University of Washington, Seattle, USA
[3]connAIx Research School for Applied AI, University of Tübingen, Heilbronn, Germany
[4]Tübingen AI Center, University of Tübingen, Tübingen, Germany
[5]Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, USA
[6]Climate/Atmospheric Science/Physical Oceanography, Scripps Institution of Oceanography, La Jolla, USA
[7]Climate Institute, Schmidt Sciences, New York, USA

Correspondence: Peter M Cox (p.m.cox@exeter.ac.uk)

Abstract. The latest Earth System Models (ESMs) are much more complex than previous generations, and are better at reproducing aspects of the current observed climate. However, in general, the latest ESMs do not reproduce the historical record of global warming with as much fidelity as the previous generation of models. We hypothesize that this is because the techniques used to calibrate ESMs have not kept track with the sophistication of the models, a phenomenon that we term the *calibration deficit*. Here we summarize the current state of the art in ESM calibration and discuss key challenges and opportunities introduced by Machine Learning and Artificial Intelligence. We outline how hybrid approaches that make use of process-based models for out-of-sample climate projections, but use ML/AI approaches for model calibration, have huge potential to revolutionize Earth System Modelling.

1 Introduction

The construction and development of Earth System Models (ESMs) has for decades followed established practices laid down by pioneers in the 1960s and 1970s. The dynamical equations (fluid flow in atmosphere and ocean) are solved through various numerical approaches for discretized Navier-Stokes on a sphere (Manabe and Wetherald, 1975). Unresolved flow, diabatic processes, other physical and chemical factors, ranging from clouds to radiative transfer to biospheric processes, are all represented in components that participate in exchanges and feedbacks with the resolved-scale fluid dynamics in atmosphere and ocean. Free parameters in components are constrained within uncertainty bounds imposed by theory or observation, and the full Earth system is then further tuned to respect observed planetary scale balances (Balaji et al., 2022).

The latest ESMs are much more complex than previous generations, as modelers aim to simulate key processes in the climate-carbon cycle system, such as cloud feedbacks and nutrient limitations on carbon sinks (see Figure 1). These models are also better at reproducing aspects of the current observed climate (Eyring et al., 2023). However, in general, the latest CMIP6 ESMs are less effective at simulating the historical record of global warming than the previous CMIP5 generation (Nijsse et al., 2020;

https://doi.org/10.5194/egusphere-2026-6437
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Tokarska et al., 2020), because many of the CMIP6 models now have climate sensitivities which are beyond the 'likely' range assessed from multiple lines of evidence (**?**). How can ESMs have much improved process representation and yet become worse at simulating the record of global warming? We hypothesize that this is largely because the techniques used to calibrate ESMs have not kept track with the sophistication of the models, a phenomenon that we term *calibration deficit*. As ESMs represent more processes, this inevitably leads to additional uncertain internal parameters, which can lead to an increase in projection uncertainty.

The process of tuning is currently laborious, based on computationally expensive and yet, too sparse exploration of uncertainty, and are often poorly documented. However, recent advances inspired by the advent of machine learning are transforming the field. A chief feature of the transformation is the passage *from simulation to emulation*: new approaches to data-model fusion allow us to calibrate both components and whole models against well-defined cost functions (Hourdin et al., 2023b; Bonnet et al., 2025a), or to enable online, derivative-free optimization that fuses observations and process-based models directly during simulation (Grundner et al., 2025). In addition, lightweight emulators allow massive exploration of parametric and structural uncertainty (Schneider et al., 2017).

The article summarizes the current state of the art in ESM benchmarking and calibration. We discuss the existing obstacles to the calibration of ESMs and their components, and outline how ML/AI could be used to rapidly advance the calibration of process-based models for out-of-sample climate projections (Price et al., 2024).

2 The Opportunities and Challenges of Machine Learning

Machine learning and Artificial Intelligence approaches are already challenging process-based Numerical Weather Prediction (Bauer et al., 2015). A number of ML/AI systems are now on a par or outperforming the state-of-the-art ECMWF forecast system, using neural networks (e.g. Pangu-Weather, Bi et al. (2023); GenCast, Price et al. (2024)). Currently, these approaches rely on a reanalysis product (e.g. ERA5) for training, and therefore remain dependent on process-based models to produce the reanalysis. However, systems are now being developed to "replace the entire NWP pipeline", which would break the dependence on process-based modelling altogether (Allen et al., 2025). Reduced-form ML-based climate models have also been developed to simulate the current climate and can now run stably for simulation periods of 1000 years (Cresswell-Clay et al., 2025).

Most climate scientists argue that purely data-driven ML/AI approaches cannot be trusted out of sample, and therefore that process-based climate models are going to be required for the foreseeable future ("in order to have confidence in projections using any modeling approach requires an underpinning of fundamental understanding", Bordoni et al. (2025)). Nevertheless, data-driven physics-constrained ML/AI models are approaching quantitative accuracy in extreme out-of-sample tests. Figure 2 shows a classic dynamical test in which the OpenIFS dynamical model is used to simulate a growing mid-latitude cyclone in a moist baroclinically unstable flow on an aqua planet. Also shown in Figure 2 is a simulation from an AI/ML model trained only on ERA5 data with observed atmospheric states over the Earth's complex distribution of land and ocean. Notably, the

https://doi.org/10.5194/egusphere-2026-6437
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

AI/ML model is trained to satisfy hydrostatic balance in each column with higher accuracy than that in the underlying ERA5 constant-pressure-level data. The zero-shot LM/AI aqua-planet result nearly matches the OpenIFS solution.

AI and ML can automatically learn subgrid physics by mapping resolved variables to missing subgrid processes. These models are trained by minimizing a cost function that quantifies the error between the predicted subgrid quantities and true values from high-resolution or observational data. While powerful, a primary drawback is that ML algorithms may violate physical laws (like conservation principles) unless specific constraints are explicitly imposed (Zanna and Bolton, 2020; Beucler et al., 2021) or larger datasets are used (Guan et al., 2022). Autoregressive AI emulators—trained on traditional models and reanalysis—offer a computationally cheap way to accelerate both short-term weather and long-term climate simulations. They enhance traditional modeling pipelines in four key ways.

1. Large Ensembles: Their low computational cost enables vastly larger ensemble sizes, improving probabilistic forecasts of climate variability.
2. Sensitivity Analysis: They allow repeated, high-fidelity simulations for highly efficient parameter tuning.
3. Flexible Boundary Forcings: Their speed makes them ideal for driving coupled simulations, such as forcing ocean models with atmospheric AI ensembles or pairing ML for slow dynamics with traditional models for fast phenomena.
4. Differentiability: Because they are inherently differentiable, they allow for gradient-based optimization and adjoint sensitivity analyses, unlocking deeper scientific insights and better tuning.

3 ESM Calibration to date

We might reasonably ask why ESM calibration is not already better than it currently is. In our view, the reasons fall into two main categories, one cultural and one practical, and these lead together to the *calibration deficit*. The first set of reasons relate to the attitude that climate modellers have traditionally had towards calibration. This arises from the historical use of climate models to demonstrate that humankind is responsible for global warming, and the related technical area of 'Detection & Attribution' (Hegerl and Zwiers, 2011). In order for climate models to be used in this way, it has been important for them not to have been calibrated against records of global warming. Although ESM components are often calibrated against available observations, it has therefore remained somewhat of a taboo to consider whole model calibration against observations. Where whole-model calibration has occurred, it has typically involved tuning to ensure that models are near a top-of-atmosphere (TOA) energy balance in their pre-industrial control climates. Tuning of this type is rarely fully documented, and partly as a result, calibration practices still vary widely across modeling centres.

The second set of reasons relate to the practical challenges involved in ESM calibration. Although some ESM components can be calibrated off-line against observations, this leads to the modeled climates drifting to unrealistic states once the components are coupled. Some ESM components (e.g. ocean heat uptake and biogeochemistry; soil and vegetation carbon) also have long intrinsic timescales, making calibration difficult because equilibrated simulations take centuries to millennia. The drive towards increasing complexity and spatial resolution, adds to this challenge by making models more computational expensive.

https://doi.org/10.5194/egusphere-2026-6437
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

This means that it is impossible to run many simulations of state-of-the-art model ESMs, as would be required for the most straightforward calibration approaches. The progress that has been made on Physics Parameter Ensembles (PPEs) has instead been achieved through using models that are faster than the latest ESMs, often models of a previous generation (Booth et al., 2012). PPE approaches are invaluable for estimating the uncertainties in climate projections (Murphy et al., 2004), but they therefore cannot be used for calibration of state-of-the-art ESMs. Where a faster older generation models has been calibrated to contemporary observations, the calibrated older generation model has been found to outperform the latest model, and the range of feasible climate sensitivities has also been shown to reduce within the calibrated model (Tett et al., 2022).

We believe that it is now critically important to develop the techniques required to calibrate state-of-the-art ESMs in order to overcome the current calibration deficit. There is still no consensus within the climate modelling community as to whether calibration should include trends, or just the statistics of the present day climate. Some argue that overfitting to the historical trends could lead to less reliable projections, because of compensating errors in different model components, while others argue that fitting to the recent rate of global warming would constrain climate sensitivities (as implied by emergent constraints such as those presented in Jiménez-de-la Cuesta and Mauritsen (2019) and Nijsse et al. (2020)). However, there is general agreement that calibration of ESMs urgently needs to become more routine and transparent, in order to make these models more useful to guide climate mitigation and adaptation over the 21^{st} and 22^{nd} centuries. As Balaji et al. (2022) put it, 'calibration, far from being a weakness of models, is an essential element in the simulation of complex systems'.

4 Evaluation versus calibration

Evaluation and calibration serve complementary but conceptually distinct roles in the Earth System modelling workflow. Model evaluation is a diagnostic and largely post hoc activity that assesses how well a model reproduces observed climate statistics and variability, using independent reference datasets, without modifying model parameters or structure. Its primary purpose is to characterize model skill, identify systematic biases, and build confidence in model behaviour across scales and applications. Model calibration, by contrast, is an interventionist process in which model parameters or parameterizations are actively adjusted to improve agreement with selected observational targets, typically guided by explicit objective functions, constraints, and assumptions about uncertainty. Maintaining a clear conceptual separation between the two is essential to avoid overfitting, to preserve the interpretability of evaluation results, and to ensure that apparent performance gains reflect genuine improvements in model realism rather than compensating errors.

Earth System Model evaluation is a central component of model development and intercomparison, providing a systematic basis for assessing model fidelity, identifying strengths and deficiencies, and guiding targeted improvements across regional, and global scales. This evaluation is typically performed by confronting model output with a broad range of observational, reanalysis, and satellite-based reference datasets using physically motivated diagnostics, performance metrics, and emergent constraints. Community tools such as the Earth System Model Evaluation Tool (ESMValTool, Eyring et al. (2020) and International Land Model Benchmarking (ILAMB, Collier et al. (2018) have established standardized, transparent, and reproducible frameworks for such analyses, enabling consistent evaluation across models, experiments, and generations, while supporting

https://doi.org/10.5194/egusphere-2026-6437
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

both global benchmarks and process-oriented diagnostics. More recently, the CMIP Rapid Evaluation Framework (Hoffman et al., 2025) has been developed to extend these capabilities toward near-real-time automated assessment of CMIP simulations as they become available, facilitating early detection of systematic biases, rapid feedback to modelling centres, and more efficient iteration between model development, evaluation, and coordinated intercomparison activities. More recently, these evaluation capabilities have been expanded to directly support model development on the native grids of Earth system models, enabling better diagnostic comparisons (Schlund et al., 2023).

A critical methodological principle is that models should not be evaluated against the same metrics, diagnostics, or observational constraints to which they were explicitly tuned or calibrated, as this risks circular validation and overestimation of true predictive skill (Rosenblatt et al., 2024). To ensure scientific integrity, modelling centers must establish a clear distinction between calibration targets and evaluation benchmarks prior to simulation. Whether a specific phenomenon, such as the El Niño–Southern Oscillation (ENSO) or the Madden-Julian Oscillation (MJO), serves as a target for tuning or an independent benchmark for skill should be determined by the model's intended use or the specific hypothesis being tested. The primary challenge in refining Earth System Models therefore lies in the strategic selection of data: one must determine which metrics will serve as calibration targets to adjust the model and which must be reserved as independent benchmarks for evaluation.

Historically, the climate modeling community has struggled to reach a consensus on how to weight individual models within multi-model ensembles based on performance, often reverting to "model democracy" where every model is treated as equally skillful. The difficulties inherent in model evaluation present a significant hurdle for benchmarking and weighting (Eyring et al., 2019), as well as for the calibration process itself. This challenge extends to machine learning emulators, which similarly require robust benchmarks to validate their performance (Watson-Parris, 2021).

Having chosen observations to calibrate against, estimates of uncertainty are not always easily available for observational datasets, with a risk that models overfit to measurement noise. If ESM calibration advances significantly as proposed in this paper, the choice of metrics to assess model quality will be key in determining what a calibrated model is good for, and what it may not be suitable for. For example, a model calibrated to fit climatological means may do so at the cost of the representation of climate variability and extremes. This would of course be counterproductive, as variability will most likely dominate the impacts of climate and climate change. Therefore, including the most vital features in the cost function is paramount and may require an iterative process. In general, there is a need to establish calibration workflows which allow recalibration to be routinely and cheaply carried-out in response to new observational data (Mai, 2023).

Calibration requires curated, standardized, and well-documented observations. Benchmarks must account for substantial internal variability, and evaluate distributions rather than single numbers. In addition, benchmark data should ideally pass foundational physical checks, such as global energy and water balance. Where observational timeseries are noisy, using multiple independent datasets with documented provenance and uncertainties, may be required to produce useful benchmarks. In order to avoid overfitting to the mean state at the cost of accurately simulating modes of variability, metrics should include variance, percentiles/extremes, return times, lag/causal relations, and cross-variable coherence (e.g., moisture–temperature coupling). Work on defining such benchmark datasets is a vital first step to enable more efficient and transparent ESM calibration.

https://doi.org/10.5194/egusphere-2026-6437
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Transparency in calibration requires moving beyond simple pass/fail metrics toward reporting performance as graded trade-offs, such as Pareto fronts or radar plots. Modelling centres (or the community) should establish minimal 'green-flag' thresholds—for example, a maximum allowable energy imbalance over a set period to provide a baseline for model reliability. Furthermore, allowing goal-specific calibration for particular policy or process needs can make models more decision-relevant, provided the risks of overfitting and the resulting trade-offs are explicitly documented so that the model spread can be interpreted rather than merely homogenized.

Machine Learning approaches routinely split observations into 'train' (i.e. calibration) and 'test' (i.e. evaluation) data, to reduce the risk of overfitting (Bishop, 1995), and reproducible calibration for ESMs will need to do likewise. An increased focus on ESM calibration will therefore require further effort to increase the observations available for calibration, and separately, for evaluation. However, we believe that the promise of faster and more transparent calibration of ESMs, which includes significantly reducing climate projection uncertainty will fully justify this effort.

In this context it is to be noted that much of our understanding of the climate and climate change come from what we might call "counterfactual" climates. For example, the most fundamental result of Manabe and Wetherald (1975) (that resulted in the 2021 Nobel prize in Physics for Syukuro Manabe) involves computing the equilibrium response of an idealized atmosphere to CO_2 increase, a condition that is never realized in nature. Detection and attribution of climate change also involves comparison to pre-industrial equilibrium climates, another counterfactual. Our confidence that calibrated models respond correctly to perturbations will very much depend on their ability to simulate or emulate counterfactual climates. Many methods in the ML literature could plausibly lead to the discovery of low-order manifolds using Variational Autoencoders (VAEs), representation learning and equation discovery. We believe that transparent calibration actually allows us to put theory and hypothesis-driven research back in the centre of the scientific method with emulators: not only do we not want to overfit, we also don't want simply to *fit*, we want to discover!

5 New Opportunities

The emerging era of ML presents an opportunity to bridge the gap between process-based, deductive understanding and observational reality. Process-based ESMs are grounded in fundamental physical principles—conservation of energy, water, and momentum—which allows them to simulate climates that have no analog in the observational record. This capability is crucial because the central challenges of climate prediction are out-of-distribution problems where data to inform relevant processes are sparse. For example, predicting how clouds will respond to rising greenhouse gases, disentangling the cooling effects of aerosols from greenhouse warming, or forecasting the future capacity of natural land carbon sinks requires extrapolating beyond what inductive inference from historical data and their embedded correlations enables. Purely data-driven approaches act as a flexible compression of the historical record; while efficient at interpolation, they lack a principled basis for extrapolation into genuinely new regimes. This limitation is particularly acute where confounding processes act simultaneously in the present climate but must be disentangled to predict future changes. This is the case, for example, in distinguishing cloud responses to surface warming from their rapid adjustments to radiative fluxes.

https://doi.org/10.5194/egusphere-2026-6437
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Conversely, process-based models rely on subgrid-scale parameterizations that are inherently imperfect and contain uncertain internal parameters. In practice, this generates the *calibration deficit*: despite being designed for out-of-sample generalization, these models often fail to capture aspects of observed variability and change (Tokarska et al., 2020; Nijsse et al., 2020; Schneider et al., 2021). The "process-based vs. data-driven" framing is therefore secondary to a more fundamental observation: calibration is, at heart, learning from noisy data in the presence of model-data mismatch. The era of ML makes it possible to recognize and operationalise formal calibration procedures within modelling centers, enabling the identification and reduction of model errors to enhance the reliability of projections.

However, to be effective, calibration must be integrated directly into the modeller's workflow, making the sources of model-data mismatch explicit and open to scrutiny. To illustrate, consider a simulator g that maps parameters $\boldsymbol{\theta}$ and initial conditions $\boldsymbol{x}_0$ to diagnostics comparable to a vector of observations $\boldsymbol{y}$, which may include higher-order statistics such as extreme values or correlation structures that encode "emergent constraints" (Schneider et al., 2017; Hall et al., 2019). A rigorous comparison requires an explicit decomposition of noise and mismatch,

$$\boldsymbol{y} = g(\boldsymbol{\theta}, \boldsymbol{x}_0; \boldsymbol{\delta}_I) + \boldsymbol{\delta}_E + \boldsymbol{\epsilon} + \boldsymbol{\eta},$$

where $\boldsymbol{\delta}_{I/E}$ represent internal or external structural discrepancies, $\boldsymbol{\epsilon}$ aggregates internal variability (e.g., from initial-condition uncertainty), and $\boldsymbol{\eta}$ represents observational uncertainty. Calibration only becomes well-defined when the behaviour of these terms is specified—a task complicated by the curse of high-dimensionality. Parameter spaces typically exceed $\mathcal{O}(100)$ dimensions, particularly when ML components are integrated, while diagnostic vectors can reach $\mathcal{O}(10^6)$ dimensions with complex correlations that are often only indirectly informative about the small-scale processes dominating uncertainties (e.g., in-cloud turbulence). Furthermore, high computational costs restrict feasible ensembles to $\mathcal{O}(100)$ forward simulations. Under these constraints, calibration is at its core a problem of experimental design: where to place a budget of $\mathcal{O}(100)$ expensive simulations in a parameter space of $\mathcal{O}(100)$ dimensions so that they are maximally informative about the parameters (Cleary et al., 2020). Classical space-filling designs, such as Latin hypercube sampling, cannot solve this problem: the number of simulations needed to cover parameter space at fixed resolution grows exponentially with its dimension, while the region consistent with observations occupies a small fraction of it, so nearly all simulations in a data-blind design will be wasted on parameter choices inconsistent with observations. Workable approaches must instead rely on explicit mismatch assumptions to allocate the limited simulation budget adaptively, exploiting the low effective dimensionality of the parameter space where possible.

Different calibration traditions solve this experimental design problem in distinct ways, and they must also contend with the fact that the objective function itself is noisy: diagnostics estimated as finite-time statistics are affected by noise from the chaotic internal variability of the climate system ($\boldsymbol{\epsilon}$). Bayesian calibration typically channels mismatch assumptions through the likelihood function, often via an assumed covariance matrix Γ. To make Bayesian inference tractable when the simulator $g(\cdot)$ is computationally expensive, the "calibrate-emulate-sample" (CES) framework (Cleary et al., 2020; Dunbar et al., 2021) breaks the problem into three steps. The calibration step uses ensemble Kalman inversion or sampling (Iglesias et al., 2013; Schillings and Stuart, 2017) as derivative-free optimizers that tolerate noisy objectives and iteratively evolve an ensemble of simulations toward regions of high posterior probability. This step solves the experimental design problem: in the continuous limit, the

https://doi.org/10.5194/egusphere-2026-6437
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

ensemble approximates a Langevin-type diffusion whose invariant measure is the posterior, so the parameter–diagnostic pairs accumulated as the ensemble evolves concentrate in regions of high posterior probability (Garbuno-Inigo et al., 2020; Cleary et al., 2020), where an emulator must be accurate for inference. The design points need not be distributed according to the posterior; it suffices that they cover the region of significant posterior mass, since the subsequent sampling step corrects the distribution. In the emulation and sampling steps, emulators (e.g., Gaussian processes or neural networks) trained on these pairs replace the expensive simulator within Markov chain Monte Carlo (MCMC) sampling of the posterior, as in simulation-based inference (Cranmer et al., 2020). This three-step process allows the method to exploit effective low dimensionality of the problem, focusing emulation efforts where they are most valuable. It inherits the scalability of ensemble Kalman methods with parameter space dimension, delivering Bayesian uncertainty quantification in $\mathcal{O}(100)$-dimensional parameter spaces with $\mathcal{O}(100)$ forward simulations (Kovachki and Stuart, 2019; Christopoulos et al., 2024; Gjini et al., 2026). Autodifferentiable software (e.g., , PyTorch, and Julia's autodiff ecosystem) offers an additional advantage. The noise from chaotic internal variability that these methods absorb renders exact simulator derivatives unusable for parameter inference from statistics (Gjini et al., 2026). But emulators can be imbued with guaranteed derivative smoothness, and access to stable gradients can accelerate sampling via gradient-based methods such as Hamiltonian Monte Carlo or Barker proposals (Duane et al., 1987; Livingstone and Zanella, 2022).

History matching (HM) solves the same design problem with a different emphasis: rather than committing to a full likelihood, HM uses emulation and uncertainty to rule out regions of parameter space that are implausible given observations and stated tolerances (Williamson et al., 2013). HM likewise proceeds by sequential design, using early ensembles for broad, inexpensive screening before concentrating later waves of evaluation in regions of low misfit. This approach has been successfully adapted for high-dimensional structured outputs (Salter et al., 2024) and deployed across a growing range of climate components (Hourdin et al., 2020; Baker et al., 2022; Hourdin et al., 2023a; Bonnet et al., 2025b; Raoult et al., 2025). HM addresses dimensionality constraints in part by employing lower-fidelity emulators to reduce the parameter space early—analogous to trust regions in Bayesian optimization (Eriksson and Jankowiak, 2021). However, scaling to very high dimensions increasingly relies on integrating low-dimensional embedding techniques from the ML literature, which require large volumes of data for training (Eriksson and Jankowiak, 2021; Papenmeier et al., 2022; Rumsey et al., 2023).

The broader adoption of ML may also address the bottleneck of error identification. The efficacy of calibration is bounded by the fidelity of the mismatch representation. Inferring this representation purely from data proves difficult without strong prior information (Kennedy and O'Hagan, 2001; Brynjarsdóttir and O'Hagan, 2014). Furthermore, treating structural error as an external output bias ($\boldsymbol{\delta}_E$) can violate physical constraints and compromise out-of-sample robustness. A more powerful approach embeds structural errors *internally* ($\boldsymbol{\delta}_I$), where the errors are actually made, as the source of mismatch (Williamson et al., 2014; Wu et al., 2024; Verma et al., 2024). Flexible ML components, such as neural networks or symbolic regression, can be learned within closure schemes to correct physics while enforcing constraints such as energy conservation. However, these internal corrections introduce their own hyperparameters, which effectively encode the magnitude of error attributed to the learned component; these hyperparameters thus become an integral part of the uncertainty specification that must be defended (McKinley et al., 2025).

https://doi.org/10.5194/egusphere-2026-6437
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

The experimental design perspective extends beyond calibration to how ensembles of simulations are planned in the first place, for example, to quantify uncertainties in climate projections. The response to the *calibration deficit* should not be to construct ever larger static perturbed-parameter, initial-condition, or multi-model ensembles: a static perturbed-parameter ensemble is analogous to a space-filling design, with the same poor scaling in high-dimensional parameter spaces. The scale of CMIP6—tens of thousands of simulated years, petabytes of data, and a substantial carbon footprint (Acosta et al., 2024)—makes the efficiency question essential if calibration ensembles are to become routine. CES and HM already present alternatives: simulations are chosen adaptively, selecting the next simulations where they are expected to be most informative. This active-learning logic is routine in Bayesian optimization and machine learning (MacKay, 1992; Settles, 2009; Gramacy, 2020; Shahriari et al., 2016; Frazier, 2018) and is built into widely used emulator software (Gramacy, 2016; Binois and Gramacy, 2021; Ming et al., 2023). Applied at scale to climate simulations, it treats emulators not as post-processing tools fitted after an ensemble is complete, but as active statistical objects that steer the simulation ensemble: short, multi-fidelity runs screen out unstable or drifting regions of parameter space; targeted long simulations are reserved for regions where uncertainty is greatest; and MIP-style questions could increasingly be answered by actively trained emulators rather than by each centre repeating large suites of expensive simulations. Recent climate model tuning using ML emulators and history matching already points in this direction (Bonnet et al., 2025a). Realizing these gains requires supercomputing allocations and workflow systems organized around adaptive ensemble building rather than large fixed batches—fewer, well-chosen simulations, targeted at reducing uncertainty.

Finally, a significant workflow challenge remains. While calibration assumes a well-defined target, ESM development often treats the model as a fluid object, where equations and diagnostics are continually altered in response to mismatches. Unless diagnostics, mismatch sources, and observation operators are co-designed with modellers to establish stable targets, no amount of statistical sophistication will deliver meaningful calibration. However, if the community invests in making mismatch explicit and interpretable, modern emulation, ensemble calibration methods, and structured internal discrepancy modeling can transform calibration into a routine component of model development. This operational shift would move model development from multi-year cycles of adhoc manual tuning to calibration-driven development executable in weeks.

6 Conclusions

Climate models have become progressively more complex through the generations CMIP3, CMIP5, and CMIP6. Models now represent more processes, such as an interactive carbon cycle, and have been relabeled Earth System Models (ESMs). In general, the latest models better reproduce the current climate. However, the CMIP6 models collectively also reproduce historical records of global warming less well than the previous CMIP5. We believe that this is in large part due to the failure of model calibration to keep pace with increasing model complexity, which leads to a *calibration deficit* in model performance. The rapid emergence of AI/ML approaches in weather forecasting poses a challenge to process-based climate modeling, but is also a huge opportunity. We recognize the need to continue to develop and use process-based models for climate projections, as these can be used out of sample to simulate ‘no analogue’ climates, and can be interrogated to understand causal mechanisms.

https://doi.org/10.5194/egusphere-2026-6437
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



However, these models need to be much better at reproducing historical climate change. On the other hand, ML/AI approaches, such as whole model emulation, offer a means to hugely improve the calibration of process-based Earth System Models, but need to be plausible under counterfactual perturbations and in idealized configurations without an observational analog. Therefore, from both the process-based and data-driven approaches, formal transparent calibration appears as a critical element in the scientific method.

Our key recommendations can be summarized as:

- Calibration needs to be systematic, transparent, and purpose-driven. Progress depends on creating shared, reproducible workflows, community datasets, and clearly defined calibration targets versus validation benchmarks.
- To deal with coupled model complexity, we need multi-scale calibration frameworks that address interactions between the atmosphere, ocean, land, and carbon cycle. Iterative coupling, model parsimony/parameter sparsity, and careful treatment of long-timescale processes are key.
- Emerging techniques such as ensemble methods, model emulation and reinforcement learning have great potential to improve ESM calibration. However, they must remain physics-informed and transparent to avoid overfitting and loss of interpretability. Building AI-ready, modular infrastructures will be essential to scale these approaches. Emerging methods such as active learning will reduce the computational burden of training by an adaptive approach to uncertainty sampling across ensembles.
- Evaluation and validation must move beyond a narrow set of metrics toward uncertainty-aware, multi-tiered assessment.
- Recent developments in both process-based and data-based models offer a huge opportunity to significantly improve ESM calibration, with the promise of making ESMs more interpretable, comparable, and decision-relevant (see Figure 3).

Author contributions. P.M.C. and V.B. planned the paper. All co-authors provided text, insight and comments.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Acknowledgements. This paper emerged from a workshop on climate model calibration in Washington D.C., October 29-31 2025, which was supported by the Schmidt Sciences 'Virtual Earth Systems Research Institute (VESRI)'.

https://doi.org/10.5194/egusphere-2026-6437
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

References

Acosta, M. C., Palomas, S., Paronuzzi Ticco, S. V., Utrera, G., Biercamp, J., Bretonniere, P.-A., Budich, R., Castrillo, M., Caubel, A., Doblas-Reyes, F., Epicoco, I., Fladrich, U., Joussaume, S., Kumar Gupta, A., Lawrence, B., Le Sager, P., Lister, G., Moine, M.-P., Rioual, J.-C., Valcke, S., Zadeh, N., and Balaji, V.: The computational and energy cost of simulation and storage for climate science: lessons from CMIP6, Geoscientific Model Development, 17, 3081–3098, https://doi.org/10.5194/gmd-17-3081-2024, 2024.

Allen, A., Markou, S., Tebbutt, W., Requeima, J., Bruinsma, W. P., Andersson, T. R., Herzog, M., Lane, N. D., Chantry, M., Hosking, J. S., et al.: End-to-end data-driven weather prediction, Nature, 641, 1172–1179, https://doi.org/10.1038/s41586-025-08897-0, 2025.

Baker, E., Harper, A. B., Williamson, D., and Challenor, P.: Emulation of high-resolution land surface models using sparse Gaussian processes with application to JULES, Geoscientific Model Development, 15, 1913–1929, https://doi.org/10.5194/gmd-15-1913-2022, 2022.

Balaji, V., Couvreux, F., Deshayes, J., Gautrais A, J., Hourdin, F., and Rio, C.: Are general circulation models obsolete?, Proceedings of the National Academy of Sciences, 119, e2202075 119–e2202075 119, https://doi.org/10.1073/pnas.2202075119, 2022.

Bauer, P., Thorpe, A., and Brunet, G.: The quiet revolution of numerical weather prediction, Nature, 525, 47–55, https://doi.org/10.1038/nature14956, 2015.

Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P., and Gentine, P.: Enforcing Analytic Constraints in Neural Networks Emulating Physical Systems, Phys. Rev. Lett., 126, 098 302, https://doi.org/10.1103/PhysRevLett.126.098302, 2021.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Chuan Tian, Q.: Accurate medium-range global weather forecasting with 3D neural networks, Nature, 619, 533–538, https://doi.org/10.1038/s41586-023-06185-3, 2023.

Binois, M. and Gramacy, R. B.: hetGP: Heteroskedastic Gaussian Process Modeling and Sequential Design in R, Journal of Statistical Software, 98, 1–44, https://doi.org/10.18637/jss.v098.i13, 2021.

Bishop, C. M.: Neural Networks for Pattern Recognition, https://doi.org/10.1093/oso/9780198538493.001.0001, 1995.

Bonnet, P., Pastori, L., Schwabe, M., Giorgetta, M., Iglesias-Suarez, F., and Eyring, V.: Tuning the ICON-A 2.6.4 climate model with machine-learning-based emulators and history matching, Geoscientific Model Development, 18, 3681–3706, https://doi.org/10.5194/gmd-18-3681-2025, 2025a.

Bonnet, P., Pastori, L., Schwabe, M., Giorgetta, M. A., Iglesias-Suarez, F., and Eyring, V.: Tuning the ICON-A 2.6.4 climate model with machine-learning-based emulators and history matching, Geoscientific model development, 18, 3681–3706, https://doi.org/10.5194/gmd-18-3681-2025, 2025b.

Booth, B. B., Jones, C. D., Collins, M., Totterdell, I. J., Cox, P. M., Sitch, S., Huntingford, C., Betts, R. A., Harris, G. R., and Lloyd, J.: High sensitivity of future global warming to land carbon cycle processes, Environmental Research Letters, 7, 024 002–024 002, https://doi.org/10.1088/1748-9326/7/2/024002, 2012.

Bordoni, S., Kang, S., Shaw, T. A., Simpson, I., and Zanna, L.: The futures of climate modeling, npj Climate and Atmospheric Science, 8, https://doi.org/10.1038/s41612-025-00955-8, 2025.

Bouvier, C., van den Broek, D., Ekblom, M., and Sinclair, V. A.: Analytical and Adaptable Initial Conditions for Dry and Moist Baroclinic Waves in the Global Hydrostatic Model OpenIFS (CY43R3), Geoscientific Model Development, 17, 2961–2986, https://doi.org/10.5194/gmd-17-2961-2024, 2024.

Brynjarsdóttir, J. and O'Hagan, A.: Learning about physical parameters: The importance of model discrepancy, Inverse Problems, 30, 114 007–114 007, https://doi.org/10.1088/0266-5611/30/11/114007, 2014.

https://doi.org/10.5194/egusphere-2026-6437
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



Christopoulos, C., Lopez-Gomez, I., Beucler, T., Cohen, Y., Kawczynski, C., Dunbar, O. R. A., and Schneider, T.: Online Learning of Entrainment Closures in a Hybrid Machine Learning Parameterization, Journal of Advances in Modeling Earth Systems, 16, https://doi.org/10.1029/2024ms004485, 2024.

Cleary, E., Garbuno-Inigo, A., Lan, S., Schneider, T., and Stuart, A. M.: Calibrate, emulate, sample, Journal of Computational Physics, 424, 109 716–109 716, https://doi.org/10.1016/j.jcp.2020.109716, 2020.

Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., Mu, M., and Randerson, J. T.: The International Land Model Benchmarking (ILAMB) System: Design, Theory, and Implementation, Journal of Advances in Modeling Earth Systems, 10, 2731–2754, https://doi.org/https://doi.org/10.1029/2018MS001354, 2018.

Cranmer, K., Brehmer, J., and Louppe, G.: The frontier of simulation-based inference, Proc. Natl. Acad. Sci., 117, 30 055–30 062, https://doi.org/10.1073/pnas.1912789117, 2020.

Cresswell-Clay, N., Liu, B., Durran, D. R., Liu, Z., Espinosa, Z. I., Moreno, R. A., and Karlbauer, M.: A deep learning earth system model for efficient simulation of the observed climate, AGU Advances, 6, https://doi.org/10.1029/2025av001706, 2025.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D.: Hybrid Monte Carlo, Physics Letters B, 195, 216–222, https://doi.org/10.1016/0370-2693(87)91197-x, 1987.

Dunbar, O. R. A., Garbuno-Inigo, A., Schneider, T., and Stuart, A. M.: Calibration and uncertainty quantification of convective parameters in an idealized GCM, J. Adv. Model. Earth Sys., 13, e2020MS002 454, https://doi.org/10.1029/2020MS002454, 2021.

Eriksson, D. and Jankowiak, M.: High-Dimensional Bayesian Optimization with Sparse Axis-Aligned Subspaces, arXiv (Cornell University), https://doi.org/10.48550/arxiv.2103.00349, 2021.

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., Collins, W. D., Gier, B. K., Hall, A. D., Hoffman, F. M., et al.: Taking climate model evaluation to the next level, Nature Climate Change, 9, 102–110, https://doi.org/10.1038/s41558-018-0355-y, 2019.

Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., Arnone, E., Bellprat, O., Brötz, B., Caron, L.-P., Carvalhais, N., Cionni, I., Cortesi, N., Crezee, B., Davin, E. L., Davini, P., Debeire, K., de Mora, L., Deser, C., Docquier, D., Earnshaw, P., Ehbrecht, C., Gier, B. K., Gonzalez-Reviriego, N., Goodman, P., Hagemann, S., Hardiman, S., Hassler, B., Hunter, A., Kadow, C., Kindermann, S., Koirala, S., Koldunov, N., Lejeune, Q., Lembo, V., Lovato, T., Lucarini, V., Massonnet, F., Müller, B., Pandde, A., Pérez-Zanón, N., Phillips, A., Predoi, V., Russell, J., Sellar, A., Serva, F., Stacke, T., Swaminathan, R., Torralba, V., Vegas-Regidor, J., von Hardenberg, J., Weigel, K., and Zimmermann, K.: Earth System Model Evaluation Tool (ESMValTool) v2.0 – an extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP, Geoscientific Model Development, 13, 3383–3438, https://doi.org/10.5194/gmd-13-3383-2020, 2020.

Eyring, V., Gillett, N. P., Achuta Rao, K. M., Barimalala, R., Barreiro Parrillo, M., Bellouin, N., Cassou, C., Durack, P. J., Kosaka, Y., McGregor, S., et al.: Human Influence on the Climate System, Climate Change 2021 - The Physical Science Basis, pp. 423–552, https://doi.org/10.1017/9781009157896.005, 2023.

Frazier, P. I.: A tutorial on Bayesian optimization, arXiv (Cornell University), https://doi.org/10.48550/arxiv.1807.02811, 2018.

Garbuno-Inigo, A., Hoffmann, F., Li, W., and Stuart, A. M.: Interacting Langevin Diffusions: Gradient Structure and Ensemble Kalman Sampler, SIAM J. Appl. Dyn. Sys., 19, 412–441, https://doi.org/10.1137/19M1251655, 2020.

Gjini, R., Morzfeld, M., Dunbar, O. R., and Schneider, T.: The Ensemble Kalman Inversion Race, Journal of Advances in Modeling Earth Systems, 18, https://doi.org/10.1029/2025ms005629, 2026.

https://doi.org/10.5194/egusphere-2026-6437
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



Gramacy, R. B.: laGP: Large-Scale Spatial Modeling via Local Approximate Gaussian Processes in R, Journal of Statistical Software, 72, 1–46, https://doi.org/10.18637/jss.v072.i01, 2016.

Gramacy, R. B.: Surrogates, Chapman and Hall/CRC, https://doi.org/10.1201/9780367815493, 2020.

Grundner, A., Beucler, T., Savre, J., Lauer, A., Schlund, M., and Eyring, V.: Reduced cloud cover errors in a hybrid AI-climate model through equation discovery and automatic tuning, Scientific Reports, 15, https://doi.org/10.1038/s41598-025-29155-3, 2025.

Guan, W., Chen, R., Zhang, H., Yang, Y., and Wei, H.: Seasonal Surface Eddy Mixing in the Kuroshio Extension: Estimation and Machine Learning Prediction, Journal of Geophysical Research: Oceans, 127, e2021JC017 967, https://doi.org/https://doi.org/10.1029/2021JC017967, e2021JC017967 2021JC017967, 2022.

Hall, A., Cox, P., Huntingford, C., and Klein, S.: Progressing emergent constraints on future climate change, Nature Climate Change, 9, 269–278, https://doi.org/10.1038/s41558-019-0436-6, 2019.

Hegerl, G. and Zwiers, F.: Use of models in detection and attribution of climate change, Wiley Interdisciplinary Reviews Climate Change, 2, 570–591, https://doi.org/10.1002/wcc.121, 2011.

Hoffman, F. M., Hassler, B., Swaminathan, R., Lewis, J., Andela, B., Collier, N., Hegedűs, D., Lee, J., Pascoe, C., Pflüger, M., Stockhause, M., Ullrich, P., Xu, M., Bock, L., Chun, F., Gier, B. K., Kelley, D. I., Lauer, A., Lenhardt, J., Schlund, M., Sreeush, M. G., Weigel, K., Blockley, E., Beadling, R., Beucher, R., Dugassa, D. D., Lembo, V., Lu, J., Brands, S., Tjiputra, J., Malinina, E., Mederios, B., Scoccimarro, E., Walton, J., Kershaw, P., Marquez, A. L., Roberts, M. J., O'Rourke, E., Dingley, E., Turner, B., Hewitt, H., and Dunne, J. P.: Rapid Evaluation Framework for the CMIP7 Assessment Fast Track, EGUsphere, 2025, 1–57, https://doi.org/10.5194/egusphere-2025-2685, 2025.

Hourdin, F., Williamson, D., Rio, C., Couvreux, F., Roehrig, R., Villefranque, N., Musat, I., Diallo, F. B., Fairhead, L., and Volodina, V.: Process-Based Climate Model Development Harnessing Machine Learning: II. Model Calibration From Single Column to Global, Journal of Advances in Modeling Earth Systems, 13, e2020MS002 225, https://doi.org/10.1029/2020ms002225, 2020.

Hourdin, F., Ferster, B., Deshayes, J., Mignot, J., Musat, I., and Williamson, D.: Toward machine-assisted tuning avoiding the underestimation of uncertainty in climate change projections, Science Advances, 9, eadf2758–eadf2758, https://doi.org/10.1126/sciadv.adf2758, 2023a.

Hourdin, F., Ferster, B., Deshayes, J., Mignot, J., Musat, I., and Williamson, D.: Toward machine-assisted tuning avoiding the underestimation of uncertainty in climate change projections, Science Advances, 9, eadf2758–eadf2758, https://doi.org/10.1126/sciadv.adf2758, 2023b.

Iglesias, M. A., Law, K. J. H., and Stuart, A. M.: Ensemble Kalman methods for inverse problems, Inverse Problems, 29, 045 001–045 001, https://doi.org/10.1088/0266-5611/29/4/045001, 2013.

Jiménez-de-la Cuesta, D. and Mauritsen, T.: Emergent constraints on Earth's transient and equilibrium response to doubled CO_2 from post-1970s global warming, Nature Geoscience, 12, 902–905, https://doi.org/10.1038/s41561-019-0463-y, 2019.

Kennedy, M. C. and O'Hagan, A.: Bayesian Calibration of Computer Models, Journal of the Royal Statistical Society Series B (Statistical Methodology), 63, 425–464, https://doi.org/10.1111/1467-9868.00294, 2001.

Kovachki, N. B. and Stuart, A. M.: Ensemble Kalman inversion: a derivative-free technique for machine learning tasks, Inverse Problems, 35, 095 005, https://doi.org/10.1088/1361-6420/ab1c3a, 2019.

Livingstone, S. and Zanella, G.: The Barker Proposal: Combining Robustness and Efficiency in Gradient-Based MCMC, Journal of the Royal Statistical Society Series B: Statistical Methodology, 84, 496–523, https://doi.org/10.1111/rssb.12482, 2022.

MacKay, D. J. C.: Information-based objective functions for active data selection, Neural Computation, 4, 590–604, https://doi.org/10.1162/neco.1992.4.4.590, 1992.

https://doi.org/10.5194/egusphere-2026-6437
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



Mai, J.: Ten strategies towards successful calibration of environmental models, Journal of Hydrology, 620, 129 414, https://doi.org/https://doi.org/10.1016/j.jhydrol.2023.129414, 2023.

Manabe, S. and Wetherald, R. T.: The Effects of Doubling the CO_2 Concentration on the climate of a General Circulation Model, Journal of the Atmospheric Sciences, 32, 3–15, https://doi.org/10.1175/1520-0469(1975)032<0003:teodtc>2.0.co;2, 1975.

McKinley, T. J., Williamson, D. B., Xiong, X., Salter, J. M., Challen, R., Danon, L., Youngman, B. D., and McNeall, D.: On real-time calibrated prediction for complex model-based decision support in pandemics: Part 2, medRxiv, https://doi.org/10.1101/2025.05.16.25327744, 2025.

Ming, D., Williamson, D., and Guillas, S.: Deep Gaussian process emulation using stochastic imputation, Technometrics, 65, 150–161, https://doi.org/10.1080/00401706.2022.2124311, 2023.

Murphy, J. M., Sexton, D. M., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., and Stainforth, D. A.: Quantification of modelling uncertainties in a large ensemble of climate change simulations, Nature, 430, 768–772, https://doi.org/10.1038/nature02771, 2004.

Nijsse, F. J., Cox, P. M., and Williamson, M. S.: Emergent constraints on transient climate response (TCR) and equilibrium climate sensitivity (ECS) from historical warming in CMIP5 and CMIP6 models, Earth System Dynamics, 11, 737–750, https://doi.org/10.5194/esd-11-737-2020, 2020.

Papenmeier, A., Kern, D., Englebienne, G., and Seifert, C.: It's complicated: The relationship between user trust, model accuracy and explanations in AI, ACM Transactions on Computer-Human Interaction, 29, 1–33, https://doi.org/10.1145/3495013, 2022.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., et al.: Probabilistic weather forecasting with machine learning, Nature, 637, 84–90, https://doi.org/10.1038/s41586-024-08252-9, 2024.

Raoult, N., Douglas, N., MacBean, N., Kolassa, J., Quaife, T., Roberts, A. G., Fisher, R., Fer, I., Bacour, C., Dagon, K., Hawkins, L., Carvalhais, N., Cooper, E., Dietze, M. C., Gentine, P., Kaminski, T., Kennedy, D., Liddy, H. M., Moore, D. J. P., Peylin, P., Pinnington, E., Sanderson, B., Scholze, M., Seiler, C., Smallman, T. L., Vergopolan, N., Viskari, T., Williams, M., and Zobitz, J.: Parameter Estimation in Land Surface Models: Challenges and Opportunities With Data Assimilation and Machine Learning, Journal of Advances in Modeling Earth Systems, 17, e2024MS004 733, https://doi.org/https://doi.org/10.1029/2024MS004733, e2024MS004733 2024MS004733, 2025.

Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S., and Scheinost, D.: Data leakage inflates prediction performance in connectome-based machine learning models, Nature Communications, 15, 1829–1829, https://doi.org/10.1038/s41467-024-46150-w, 2024.

Rumsey, K. N., Huerta, G., and Derek Tucker, J.: A localized ensemble of approximate Gaussian processes for fast sequential emulation, Stat, 12, https://doi.org/10.1002/sta4.576, 2023.

Salter, J. M., Webster, H. N., and Saint, C.: Technical note: Exploring parameter and meteorological uncertainty via emulation in volcanic ash atmospheric dispersion modelling, Atmospheric Chemistry and Physics, 24, 6251–6274, https://doi.org/10.5194/acp-24-6251-2024, 2024.

Schillings, C. and Stuart, A. M.: Analysis of the Ensemble Kalman Filter for Inverse Problems, SIAM Journal on Numerical Analysis, 55, 1264–1290, https://doi.org/10.1137/16m105959x, 2017.

Schlund, M., Hassler, B., Lauer, A., Andela, B., Jöckel, P., Kazeroni, R., Loosveldt Tomas, S., Medeiros, B., Predoi, V., Sénési, S., Servonnat, J., Stacke, T., Vegas-Regidor, J., Zimmermann, K., and Eyring, V.: Evaluation of native Earth system model output with ESMValTool v2.6.0, Geoscientific Model Development, 16, 315–333, https://doi.org/10.5194/gmd-16-315-2023, 2023.

Schneider, T., Lan, S., Stuart, A., and Teixeira, J.: Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations, Geophysical Research Letters, 44, https://doi.org/10.1002/2017gl076101, 2017.

https://doi.org/10.5194/egusphere-2026-6437
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



Schneider, T., Jeevanjee, N., and Socolow, R.: Accelerating progress in climate science, Physics Today, 74, 44–51, https://doi.org/10.1063/pt.3.4772, 2021.

Settles, B.: Active Learning Literature Survey, Computer Sciences Technical Report 1648, University of Wisconsin-Madison, https://doi.org/https://openalex.org/W2903158431, 2009.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and de Freitas, N.: Taking the human out of the loop: a review of Bayesian optimization, Proceedings of the IEEE, 104, 148–175, https://doi.org/10.1109/JPROC.2015.2494218, 2016.

Tett, S. F., Gregory, J. M., Freychet, N., Cartis, C., Mineter, M. J., and Roberts, L.: Does model calibration reduce uncertainty in climate projections?, Journal of Climate, 35, 2585–2602, https://doi.org/10.1175/jcli-d-21-0434.1, 2022.

Tokarska, K. B., Stolpe, M. B., Sippel, S., Fischer, E. M., Smith, C. J., Lehner, F., and Knutti, R.: Past warming trend constrains future warming in CMIP6 models, Science Advances, 6, eaaz9549–eaaz9549, https://doi.org/10.1126/sciadv.aaz9549, 2020.

Verma, Y., Heinonen, M., and Garg, V.: Climode: Climate and weather forecasting with physics-informed neural odes, arXiv (Cornell University), https://doi.org/10.48550/arxiv.2404.10024, 2024.

Watson-Parris, D.: Machine learning for weather and climate are worlds apart, Philosophical Transactions of the Royal Society A Mathematical Physical and Engineering Sciences, 379, 20200 098–20200 098, https://doi.org/10.1098/rsta.2020.0098, 2021.

Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., and Yamazaki, K.: History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble, Climate Dynamics, 41, 1703–1729, https://doi.org/10.1007/s00382-013-1896-4, 2013.

Williamson, D., Blaker, A. T., Hampton, C., and Salter, J.: Identifying and removing structural biases in climate models with history matching, Climate Dynamics, 45, 1299–1324, https://doi.org/10.1007/s00382-014-2378-z, 2014.

Wu, J.-L., Levine, M. E., Schneider, T., and Stuart, A. M.: Learning about structural errors in models of complex dynamical systems, Journal of Computational Physics, 513, 113 157–113 157, https://doi.org/10.1016/j.jcp.2024.113157, 2024.

Zanna, L. and Bolton, T.: Data-Driven Equation Discovery of Ocean Mesoscale Closures, Geophysical Research Letters, 47, e2020GL088 376, https://doi.org/https://doi.org/10.1029/2020GL088376, e2020GL088376 10.1029/2020GL088376, 2020.

https://doi.org/10.5194/egusphere-2026-6437
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

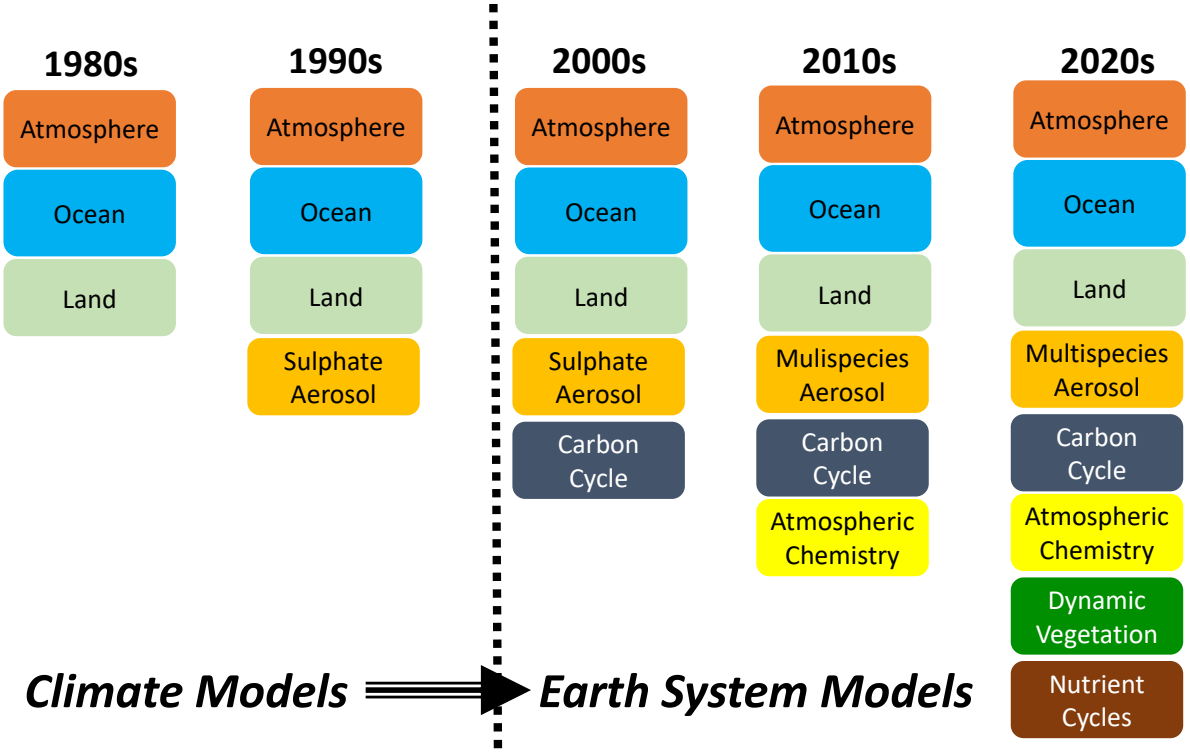


Figure 1. Schematic showing the evolution of climate models into Earth System Models.

https://doi.org/10.5194/egusphere-2026-6437
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

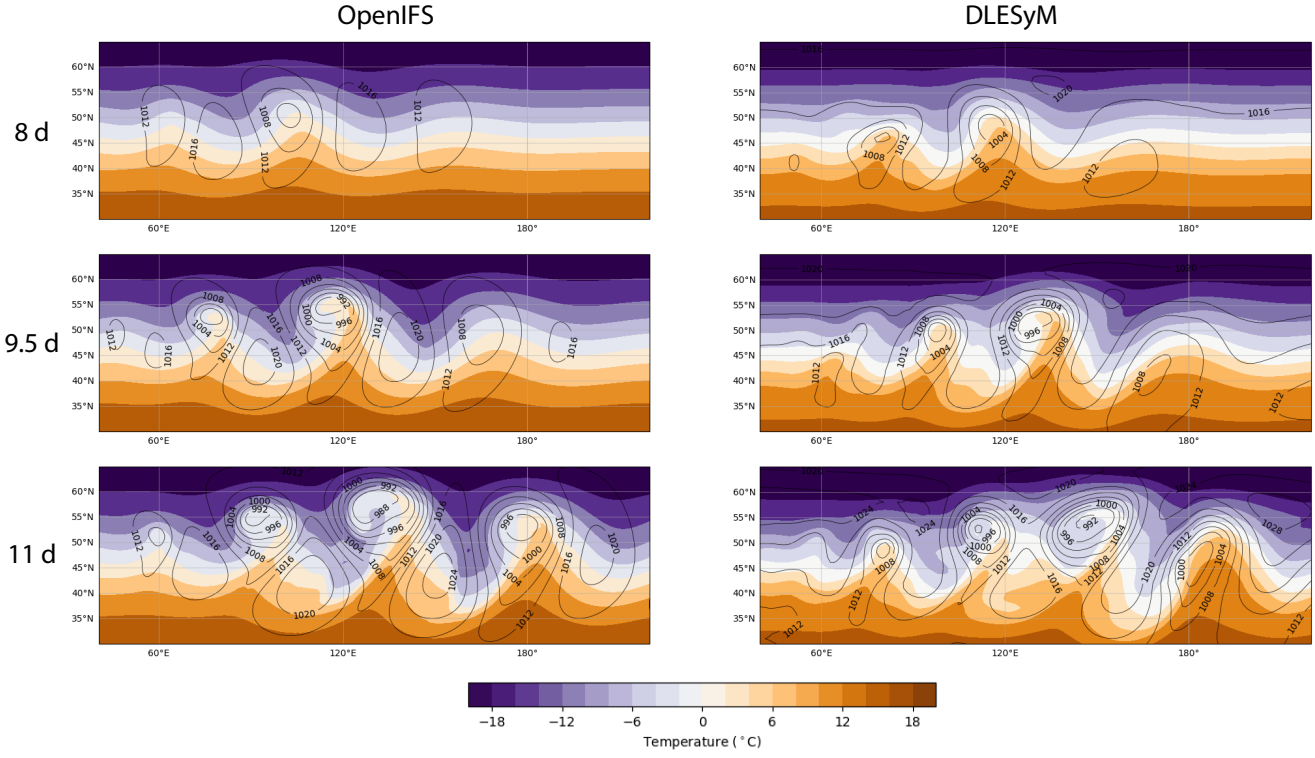


Figure 2. Comparison of OpenIFS and DLESyMv2 simulations of cyclone growth in an idealized aqua-planet test. Color fill: 2-m temperature, contours: MSLP (hPa), 8, 9.5 and 11 days after seeding the instability. Left column: dynamical solution computed using the OpenIFS from ECMWF Taken from Bouvier et al. (2024). Right column: corresponding ML/AI simulation from model trained only on ERA5 analyses, courtesy William Yik.

https://doi.org/10.5194/egusphere-2026-6437
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

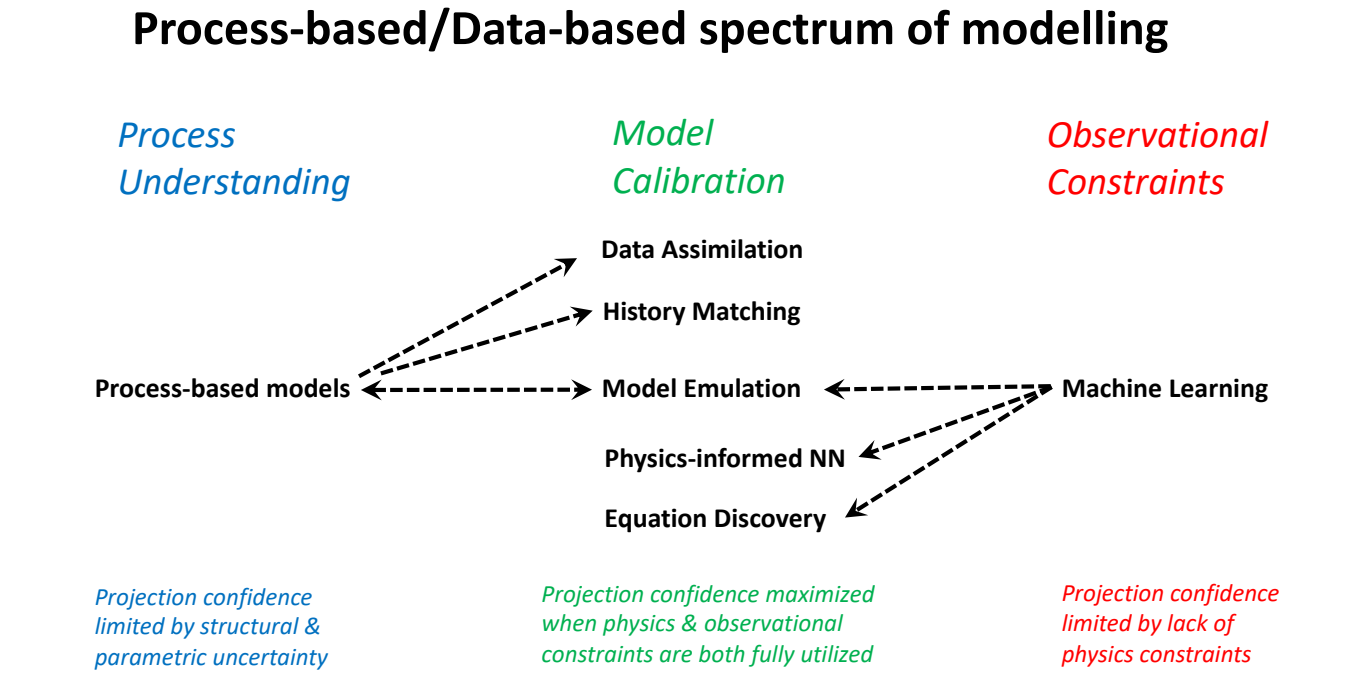


Figure 3. Schematic of the spectrum of process-based versus data-based/machine learning modeling. The calibration of process based models connects the two ends of this spectrum