

ACCEPTED MANUSCRIPT • OPEN ACCESS

Emulating the Forced Response of Climate Models with Generative Machine Learning

To cite this article before publication: Graham Clyne *et al* 2026 *Mach. Learn.: Earth* in press <https://doi.org/10.1088/3049-4753/aeac2f>

Manuscript version: Accepted Manuscript

Accepted Manuscript is “the version of the article accepted for publication including all changes made as a result of the peer review process, and which may also include the addition to the article by IOP Publishing of a header, an article ID, a cover sheet and/or an ‘Accepted Manuscript’ watermark, but excluding any other editing, typesetting or other changes made by IOP Publishing and/or its licensors”

This Accepted Manuscript is © 2026 The Author(s). Published by IOP Publishing Ltd.



As the Version of Record of this article is going to be / has been published on a gold open access basis under a CC BY 4.0 licence, this Accepted Manuscript is available for reuse under a CC BY 4.0 licence immediately.

Everyone is permitted to use all or part of the original content in this article, provided that they adhere to all the terms of the licence <https://creativecommons.org/licenses/by/4.0>

Although reasonable endeavours have been taken to obtain all necessary permissions from third parties to include their copyrighted content within this article, their full citation and copyright line may not be present in this Accepted Manuscript version. Before using any content from this article, please refer to the Version of Record on IOPscience once published for full citation and copyright details, as permissions may be required. All third party content is fully copyright protected and is not published on a gold open access basis under a CC BY licence, unless that is specifically stated in the figure caption in the Version of Record.

View the [article online](#) for updates and enhancements.

Journal Name

Crossmark

RECEIVED
dd Month yyyy

REVISED
dd Month yyyy

PAPER

Emulating the Forced Response of Climate Models with Generative Machine Learning

Graham Clyne^{1,*}, Julia Kaltenborn², Peer Nowack³, Claire Monteleoni¹, Anasatase Charantonis¹

¹INRIA, Paris, France
²MILA, Montreal, Canada ³Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
*Author to whom any correspondence should be addressed.

E-mail: clyne.graham@gmail.com

Keywords: Machine Learning, Climate Model Emulation, Future Climate, Generative Machine Learning

Abstract

Global climate models are essential tools to simulate past and potential future pathways of climate change, as well as associated climate impacts. Shared Socioeconomic Pathways (SSPs) describe a range of future scenarios of global economic and demographic development. These SSPs are intrinsically linked to changes in climate forcings—the external drivers, such as greenhouse gas and aerosol emissions, which change Earth’s energy balance over time. These forcings act as boundary conditions in Earth System Models (ESM), providing insight into the potential climatic impacts of each SSP. Running an ESM, however, is extremely computationally expensive, conflicting with the need for large ensemble runs to provide robust estimates in the presence of internal variability and scenario uncertainty. Machine Learning emulators provide a promising avenue toward fast and cheap scenario generation, but until recently lacked uncertainty quantification and the ability to condition on external forcings. Here, we build upon recent work [1] and extend it by training on multiple SSPs. We successfully generate scenarios of IPSL-CM6A-LR unseen during training and that remain largely physically consistent with the underlying climate model, including under moderate extrapolation. Under an abrupt-4×CO₂ step, however, the model does not reproduce the transient climate response. Our emulator is validated against MESMER-M, a statistical emulator of land surface temperature. Our research demonstrates that our model, ArchesClimate-SSP, does not simply imitate scenarios seen during training, but is actually capable of modeling the response of a climate state to diverse forcings. This is an important step toward reliable and rapid climate model scenario generation.

1 Introduction

Humanity is facing the consequences of a changing climate, ranging from vanishing ecosystem services (benefits that humans derive from the natural environment) to increased exposure to extreme events, putting the lives of humans, flora, and fauna at risk [2]. To move forward, we need to mitigate, adapt, and develop our future sustainably [3]. A cornerstone of climate change adaptation is humanity’s capacity to project the future of our climate under different scenarios. Policymakers need to understand how different emission plans might affect Earth’s climate to make informed decisions on behalf of the groups they represent. For decades, scientists have contributed to understanding Earth’s future climate by developing climate models [4]. Climate models started by modeling only the atmosphere and ocean, but recently they have developed into Earth System Models (ESMs), i.e. they simulate different parts of the Earth in tandem, using physical equations that formalize what we know about how the atmosphere, ocean, cryosphere and biosphere operate and interact [5, 6]. ESMs are run for many different “what-if” experiments, helping scientists better understand the Earth’s climate [4, 7]. One set of such experiments is called ScenarioMIP [8], which compares the climate outputs of different ESMs under different future scenarios. These experiments essentially ask “how will our Earth’s climate look if humanity behaves in a certain way?”. Those future scenarios of human behavior are called Shared Socioeconomic Pathways (SSPs) [9]. They describe the future in terms of economic, societal, and demographic changes over the coming decades and as such also formulate assumptions about future emissions of greenhouse

gases and aerosols. SSPs have further implications for the secondary formation of other forcing agents through atmospheric chemical reactions, in particular tropospheric and stratospheric concentrations of ozone, as well as the atmospheric oxidizing capacity (relevant for the degradation of greenhouse gases such as methane). Changes in anthropogenic greenhouse gases and aerosols expected from each SSP are provided through *Input4MIPs*, a coordinated activity that provides standardized, version-controlled datasets, providing time-dependent emission maps or atmospheric concentration fields that are then used as inputs for the ESMs participating in ScenarioMIP [10]. In summary, we can, for example, input a future emission scenario into an ESM to simulate its long-term climate response for a given SSP.

The resulting projections have several types of uncertainty. In the near-term, i.e. 30 years, model uncertainty (how unsure we are about the output of the model) is a major component of the total uncertainty, followed by the climate system’s internal variability (the natural, stochastic fluctuations within the climate system), and a small amount of scenario uncertainty (the range of potential future climate outcomes resulting from different plausible pathways of human activity). As the projection moves farther into the future, scenario uncertainty can start to become the dominant uncertainty given a wide range of SSPs [11, 12]. Unlike model uncertainty, which could be narrowed through improved representations of model physical processes and computing power, scenario uncertainty is irreducible because it pertains to future human agency. Understanding and quantifying this uncertainty is critical for policymakers; it underpins the exploration of strategies that remain effective across various trajectories. By generating different scenarios, researchers can better demonstrate how immediate policy interventions could directly alter the long-term climate. However, ESMs are computationally expensive, restricting the generation of a wide variety of future scenarios. This work aims to help climate models explore additional scenarios by providing a robust method to quickly generate spatially and temporally explicit future climate simulations that respond appropriately to a multitude of external forcings.

In this study, we investigate whether generative Deep Learning (DL) can generate unseen future climate scenarios from a set of external forcings. We also examine whether the emulator learns the correct response to each forcing provided. We use ArchesClimate as the architectural framework, an autoregressive probabilistic climate model emulator based on ArchesWeather [1, 13] and train it on a set of climate projections produced with the IPSL-CM6A-LR ESM [14]. We extend the original ArchesClimate framework by conditioning on sets of forcing agents that represent trajectories under different SSP scenarios. We will refer to the extended model generally as ArchesClimate-SSP (AC-SSP). We use monthly averages from 2015-2100 for eight future scenarios, in addition to historical runs and pre-industrial control runs. Five of these scenarios are used for training, and we hold out three SSP scenarios for evaluation: SSP4-3.4 and SSP5-8.5 from our validation set, used for architecture and hyperparameter selection, while SSP5-3.4 is a held-out test scenario never used for model selection.

We structure the paper as follows: we first provide a review of the current state of climate model emulation. We then explain the methods used in our research. Following this, we investigate several architectural and training design choices. We show that we can accurately emulate SSP scenarios unseen at training time, compared to a state-of-the-art statistical emulator, MESMER-M [15, 16]. Next, we investigate the model’s ability to learn the impact of each forcing. Finally, we apply AC-SSP to several out-of-distribution tasks, and apply our method to another climate model, CanESM5.

The contributions of this research are as follows:

- Introducing novel techniques to encourage our model to learn the relationship between the external forcings and the climate state.
- Providing a DL model that reproduces the long-term climate statistics of SSPs unseen at training time, including under moderate extrapolation.
- Showing that spatial structures of climate variability and responses can be better emulated with a training objective that adheres to spatial and temporal structures.

2 Background

To contextualize our proposed method for accelerating SSP generation, it is necessary to examine previous efforts to improve the computational efficiency of ESMs. We begin by outlining the computational bottlenecks of traditional models, followed by a review of traditional non-DL emulators. We then explore the recent paradigm shift toward machine learning approaches, concluding with the ongoing challenge of incorporating external forcings into these frameworks.

Historically, the primary alternative to ESMs has been the development of reduced-order or intermediate complexity models. These tools trade the complexity of an ESM for computational efficiency, allowing for the rapid simulation of diverse scenarios [17–19]. To generate spatially explicit data at low computational costs, these emulators are often combined with pattern scaling techniques [20–22] or functions based on historical temperature and precipitation relationships [23]. More advanced statistical emulators, such as MESMER and its monthly extension MESMER-M, successfully model internal variability and regional trends for temperature based on global yearly values [15, 16]. The most direct method for improving the speed of ESMs is to increase their computational resources. However, ESMs are already computationally expensive, and the associated engineering and operational costs remain prohibitively high [24]. While researchers are increasingly leveraging automatic differentiation and GPU computing to optimize the performance of modern physics-based climate simulations [25–27], further research is needed to represent full ESMs, driving the need for alternative modeling approaches.

Deep learning (DL) has emerged as a means to decouple computational performance from physical complexity. While DL applied to weather forecasting has already demonstrated the ability to match or exceed state-of-the-art physics-based models for short-term predictions [13, 28, 29], DL applied to climate models aims to translate these gains to decadal and centennial timescales. However, while ML approaches apply well to initial-value modeling problems, constraint-based long-term climate modeling poses novel challenges.

Architectures used to tackle these problems are categorized into deterministic models, which produce a single trajectory from a given initial condition [30, 31], and generative models, such as diffusion or flow-matching networks, which generate an ensemble of trajectories for a single state [32, 33]. Generative models create plausible samples that are particularly effective at capturing the stochastic variability inherent in climate systems. While recent models have successfully emulated both isolated ocean components and coupled ocean-atmosphere dynamics, these efforts remain largely focused on stationary climate conditions [34–36].

Despite these advancements, capturing the non-stationarity of a long-term forced climate response to external drivers like greenhouse gases and solar irradiation remains a significant hurdle. While some models utilize prescribed sea surface temperatures (SSTs) as inputs, this approach is often criticized as it avoids the fundamental task of simulating how the atmosphere responds to greenhouse gases and aerosols, as the resulting climate trajectory is already largely dictated by the input. There is a notable scarcity of ML emulators capable of prognostically modeling the Earth's response to Shared Socioeconomic Pathways (SSPs) internally rather than relying on pre-determined ocean conditions, which remains a critical gap for contributing to meaningful policy-making. Early studies include [37–40], which use long time series of emissions from forcing agents to predict climate projections for a limited set of climate variables. In contrast to early studies, more recent work includes a larger set of variables and autoregressive rollouts; however, they struggle to capture shifting distributions. For example, recent work like ACE2-SOM has explored carbon dioxide (CO₂) sensitivity at fixed warming levels (e.g., 2°C or 4°C), but these models often exhibit extreme sensitivity to distribution shifts [41, 42]. Similarly, AeroGP uses Gaussian processes to capture regional temperature future responses to changing aerosols [43]. While more approaches are beginning to attempt future state generation based on SSPs [44, 45], accurately modeling the spatiotemporal trajectory of a forced climate remains an open question that our methodology seeks to address.

3 Methods

This section introduces the dataset used in this study. We then detail the model architecture and training procedures, concluding with an overview of the metrics and analytical methods used for evaluation.

3.1 Forcings

Our model incorporates forcings from multiple sources, including greenhouse gases, aerosols and ozone. In this work, we incorporate forcings as concentrations and not as emissions. We consider learning the relationship between forcing concentrations and climate states as the first necessary step, and leave direct mappings from emissions to climate states to future work. Here we describe each forcing in detail. Please see table 1 for a summary of the forcings included in AC-SSP.

We incorporate concentrations for three well-mixed greenhouse gases (GHGs) – carbon dioxide (CO₂), methane (CH₄), and nitrous oxide (N₂O) – sourced from the *input4MIPs* repository [10, 46]. These three gases, measured in parts per billion (ppb) for CH₄ and N₂O, and parts per million (ppm) for CO₂, were selected as our primary forcing agents due to their dominant roles in

the historical anthropogenic greenhouse effect and under typical future SSPs [47]. CO₂ is the major anthropogenic greenhouse gas by historical radiative forcing and also under future SSPs, driving long-term climate change due to its longevity and large, persistent increases in emissions. Methane also poses a substantial climate forcing but has a much shorter e-folding time (the time it takes for the concentration of a gas to decrease to approximately 37% of its original amount) of around 10 years, resulting in potentially faster, scenario-dependent fluctuations in its atmospheric concentrations [47]. N₂O has an estimated e-folding time of more than 100 years and is the third most important anthropogenic greenhouse gas directly emitted into the atmosphere. These gases are well mixed, meaning they are relatively equally distributed throughout the atmosphere due to their sufficiently long lifetimes for the atmospheric circulation to distribute them across the entire atmosphere. The GHG concentrations from input4MIPS are at a 0.5-degree latitudinal monthly resolution, which we regrid to the IPSL model resolution (ca. 2.5×1.25 degrees) using first-order conservative regridding. Conservative regridding is a numerical method for transferring data between different grids that ensures the total integral of a quantity, like heat or moisture, is preserved so that no energy or mass is artificially created or lost.

Next to greenhouse gases, anthropogenic aerosol emissions pose another major human impact on Earth's energy budget, which we consider here. Such aerosols are tiny solid particles or liquid droplets suspended in the atmosphere. We use the atmospheric load for six aerosol species: Aerosol Soluble Nitrate (ASNO3M), Coarse Soluble Nitrate (CSNO3M), Coarse Insoluble Nitrate (CINO3M), Sulfate (SO₄), and two black-carbon species, Elemental Carbon Mode Accumulation Insoluble (AIBCM) and Elemental Carbon Mode Accumulation Soluble (ASBCM). In particular, sulfate aerosols are known to reflect sunlight, causing a counteracting cooling balance on the historical climate [48], including important non-linear temperature effects [45, 49, 50]. We use data from INCA, a chemistry and aerosol model developed at the Laboratoire des Sciences du Climat et de l'Environnement [51, 52]. INCA is run beforehand to generate aerosol states that are then used downstream in the IPSL model. We again conservatively regrid this data to a 144×144 resolution from 143×144 . Here we simply include the atmospheric load (kg m⁻²) of the aerosols as a rough proxy of Aerosol Optical Depth.

Finally, we differentiate ozone from the other greenhouse gases, because ozone is formed through chemical reactions in the atmosphere rather than being directly emitted from the Earth's surface [53]. In addition, its concentrations differ by orders of magnitude between the troposphere and the stratosphere. This is due to very different ozone production and loss pathways that depend on the availability of sunlight, ozone-depleting substances, and (tropospheric) ozone precursors. For example, in the troposphere, primary ozone production pathways depend on the availability of volatile organic compounds (VOCs) – including methane – and of nitrogen oxides (NO_x) from both biogenic and anthropogenic sources at Earth's surface. In the stratosphere, ozone production is primarily driven by the Chapman reactions in the presence of high-frequency ultraviolet sunlight, in balance with a variety of major catalytic ozone loss reactions. The latter also depend on anthropogenic emissions of N₂O and, famously, on a variety of halogen-containing ozone-depleting substances, which have led to the development of substantial Antarctic ozone holes over past decades [54]. In our framework, we use ozone mole fractions from *input4MIPs* covering the troposphere, stratosphere, and even the mesosphere, and conservatively regrid the data from 96×144 (latxlon) to 144×144 [55]. There are 66 vertical layers ranging from 0.0001 hPa to 1000 hPa, but we only include the following layers: [1000, 850, 500, 200, 100, 50, 10, 1, 0.01, 0.0001]. We expect ozone to help inform both stratospheric and tropospheric cooling and warming trends [56–59].

We include only a subset of the forcings provided to IPSL-CM6A-LR due to project-based computational constraints of this study and our focus on atmospheric dynamics [51]. For example, we assume that the climatic influence of ocean biogeochemical data and land-use cover data is of lower magnitude and can be omitted as a compromise with computational cost and model complexity, and left for future work.

3.2 Climate Model Data

We include monthly data from the IPSL-CM6A-LR for eight SSP scenarios, the *historical* scenario, and the *piControl* scenario. Following ScenarioMIP conventions, the scenarios range from an approximate global radiative forcing of 1.9 W m⁻² – a strong climate change mitigation scenario – to 8.5 W m⁻² by 2100. In general, we train our model on all ensemble members (which sample internal variability uncertainty) of each scenario; however, we include only 10 of the 31 available members of the historical run to keep the model focused on future scenarios. Please see Table 5 for the number of members used for each scenario, alongside a high-level description.

Forcings

Variable Name	Description	Unit
CO_2	Carbon Dioxide	ppm
CH_4	Methane	ppb
N_2O	Nitrous Oxide	ppb
$ASNO3M$	Atmospheric load of nitrate mode accumulation soluble	$kg\ m^{-2}$
$CSNO3M$	Atmospheric load of nitrate mode coarse soluble	$kg\ m^{-2}$
$CINO3M$	Atmospheric load of nitrate mode coarse insoluble	$kg\ m^{-2}$
SO_4	Atmospheric load of sulfate mode accumulation soluble	$kg\ m^{-2}$
$AIBCM$	Atmospheric load of elemental carbon mode accumulation insoluble	$kg\ m^{-2}$
$ASBCM$	Atmospheric load of elemental carbon mode accumulation soluble	$kg\ m^{-2}$
<i>Ozone</i>	Ozone mole fraction, 10 of the 66 available vertical levels (1000, 850, 500, 200, 100, 50, 10, 1, 0.01, 0.0001 hPa)	ppb

Table 1: External forcings used in AC-SSP. Each forcing is normalized and provided as an input based on its spatial dimensionality. With the exception of atmospheric aerosol loads, which are generated by the INCA chemistry and aerosol model, all forcings are sourced from *input4MIPs*. These data are generated independently of the target IPSL model and are treated as exogenous variables. Forcings are provided as inputs only for the current timestep and are not targets for autoregressive prediction. The physical justification for including each forcing is given in [subsection 3.1](#).

The variables included in AC-SSP are summarized in [Table 6](#). We select this set of variables for three reasons. First, we prioritize variables that represent physical processes governing the climate system rather than surface temperature alone, including surface fluxes (*net_surface_flux*, *evspsbl*) and a comprehensive vertical profile of atmospheric dynamics (*hus*, *ta*, *ua*, *va*, *zg* across 17 pressure levels), so that the emulator must learn the coupled thermodynamic and dynamic structure of the atmosphere rather than a single diagnostic quantity. Second, the atmospheric variable set is chosen to be comparable with those used by other AI weather and climate emulators [13, 29, 42], enabling more direct comparison of performance and behavior across the field. Third, we include an oceanic component (*thetao* across 10 depths) to capture aspects of coupled model dynamics, since IPSL-CM6A-LR is itself a coupled atmosphere-ocean model and surface climate trends are strongly modulated by ocean heat uptake and storage.

For the evaluation of our model, we utilize SSP4-3.4, abrupt-4xCO₂, and SSP5-8.5 for validation, and SSP5-3.4 for the test set. We use *historical*, *piControl*, SSP1-1.9, SSP1-2.6, SSP2-4.5, SSP3-7.0 and SSP4-6.0 as training data. SSP5-3.4 offers a valuable testing scenario due to its overshoot profile, in which CO₂ concentrations peak mid-century before declining. Overshoot dynamics are becoming increasingly likely in the future, and are therefore important to accurately model [60]. As it is our only scenario never used for architecture or hyperparameter selection, it is also our sole check against overfitting to the validation set itself. We use SSP5-8.5 and abrupt-4xCO₂ to evaluate our model’s ability to handle out-of-distribution data, a key assessment of our study. Specifically, around the year 2081 – approximately 66 years into the dataset – the CO₂ concentrations in SSP5-8.5 exceed the maximum levels seen in the training data (SSP3-7.0 giving the highest CO₂ concentration). SSP5-3.4 starts in 2040, and therefore only provides 60 years of data. SSP4-3.4 and SSP5-8.5 start in 2015 and provide 85 years of data. The inclusion of abrupt-4xCO₂ specifically allows us to assess AC-SSP’s ability to distinguish between internal climatic responses and external forced signals. As the experiment is an immediate 4x step of CO₂, it requires the model to understand the difference between the CO₂ signal and the previous climate state. We hypothesize that this experiment will show if the model is overfitting to the forcing signal.

We select the IPSL-CM6A-LR model [14] as our primary data source for several reasons. Its grid resolution ($1.25^\circ \times 2.5^\circ$) provides a computationally efficient yet physically dense dataset ideal for training deep generative emulators. Furthermore, the model’s atmospheric component, LMDZ6, incorporates advanced parameterizations for tropical convection and boundary layer dynamics, providing a state-of-the-art target for our machine learning objectives [61]. A full description of the IPSL scenario ensemble used for training and evaluation, and of the climate variables predicted by AC-SSP, is given in the Appendix ([subsection A.1](#)).

3.3 Architecture

For our deep learning framework, we employ AC-SSP, an extension of the ArchesWeather architecture [1, 13]. ArchesWeather is a Shifted Window (Swin) Vision Transformer based on the

Pangu-Weather model [62, 63]. It efficiently processes weather/climate states by partitioning spatial regions into tokens, which are then processed through successive attention blocks. These blocks utilize a windowing mechanism that shifts in location between successive layers to capture multi-scale dependencies. As opposed to the methodology of both ArchesWeather and ArchesClimate, which splits the prediction into two tasks with a deterministic and probabilistic component, we instead train the model with an Energy-score loss [64, 65]. We found that the single-step optimization of the residual of the deterministic model focused too much on short term variance, and degraded the longer forced response the deterministic model had learned. The shortcomings of the flow-matching setup are described in section A.7.

We incorporate external forcings through two distinct pathways. Spatially resolved forcings (ozone, aerosols, and spatial GHGs) are, by default, concatenated directly to the input state. In addition to channel-wise concatenation, spatially resolved forcings can be injected via a SPADE-style (SPatially-ADaptive dEnormalization) pathway [66]: a lightweight convolutional projector patch-embeds each spatial forcing field to a per-token embedding at the same resolution as the backbone’s tokens, which is then added to the token-wise shift and scale parameters of the conditioning (rather than being broadcast as a single global vector) [67].

3.3.1 Forcing Dropout To help learn the forced response, forcing channels are independently dropped at training time. Rather than substituting zeros for a dropped channel, which could otherwise be confused with a forcing that is legitimately near zero, we substitute a learned null token: a per-channel learned constant for spatial forcings. This forces the model to try to learn the role of each masked forcing by asking it to operate without this forcing at training time, so that it must infer the contribution of that forcing.

During training, spatial and non-spatial forcing channels are independently masked and replaced with the learned null token described above. Independent per-channel masking, however, means the model rarely sees every forcing at once. We therefore additionally sample, with probability p_{all} , a training step in which no forcing is masked. We evaluate different values of p_{all} to determine how often the model needs to see the full set of forcings jointly during training in order to use them correctly at inference time, when they are always fully present; $p_{\text{all}} = 0.8$ is the value used by the default model in this research. Please see section A.2 for a concrete example.

Rather than masking every forcing channel fully independently, physically coupled channels are dropped as a single unit. Two such groups are used: the six aerosol species (*ASNO3M*, *CSNO3M*, *CINO3M*, *SO4*, *AIBCM*, *ASBCM*) and the ten ozone bands described in Table 1. Every other channel (the three well-mixed greenhouse gases) is masked independently.

Formally, at each training step every channel c (or, for the two physically coupled groups above, every such group jointly) is assigned a binary keep-mask m_c by a two-stage draw:

$$r \sim \text{Uniform}(0, 1), \quad m_c = \begin{cases} 1, & r < p_{\text{all}} \\ \sim \text{Bernoulli}(1 - p_{\text{drop}}), & r \geq p_{\text{all}} \end{cases} \quad (1)$$

with r drawn once per sample (so that all channels enter the same branch, being either all kept or subject to masking) and, within the masking branch, each channel’s (or group’s) Bernoulli draw made independently.

3.4 Training and Inference

While training and inference follow the ArchesClimate framework, we implement the following modifications. First, model parameters are updated using an Exponential Moving Average (EMA) to improve stability of long autoregressive generations, and EMA weights are used exclusively for inference. Second, training proceeds in two phases. The first 16,000 steps use the Energy/Variogram loss described below, which is sufficient for the model to learn the broad structure of the climate state. Subsequently, we switch to a “pushforward” loss to learn the finer structure of the climate state. In this stage, the model is unrolled for multiple steps but gradients are retained only for the final 2 steps. This mechanism forces the model to remain accurate even after several steps of its own compounding error, which we expect to trade a degree of short-horizon precision for long-horizon and extrapolation robustness. We evaluate lengths of $\{2, 4, 8\}$ steps on our validation scenario, SSP4-3.4 (see subsection 4.1.1), and default to 4 for all subsequent experiments.

3.4.1 Energy Score Loss Here we introduce the energy score [65, 68], a proper scoring rule that rewards both accuracy and calibrated dispersion. Unlike the traditional per-variable CRPS, which

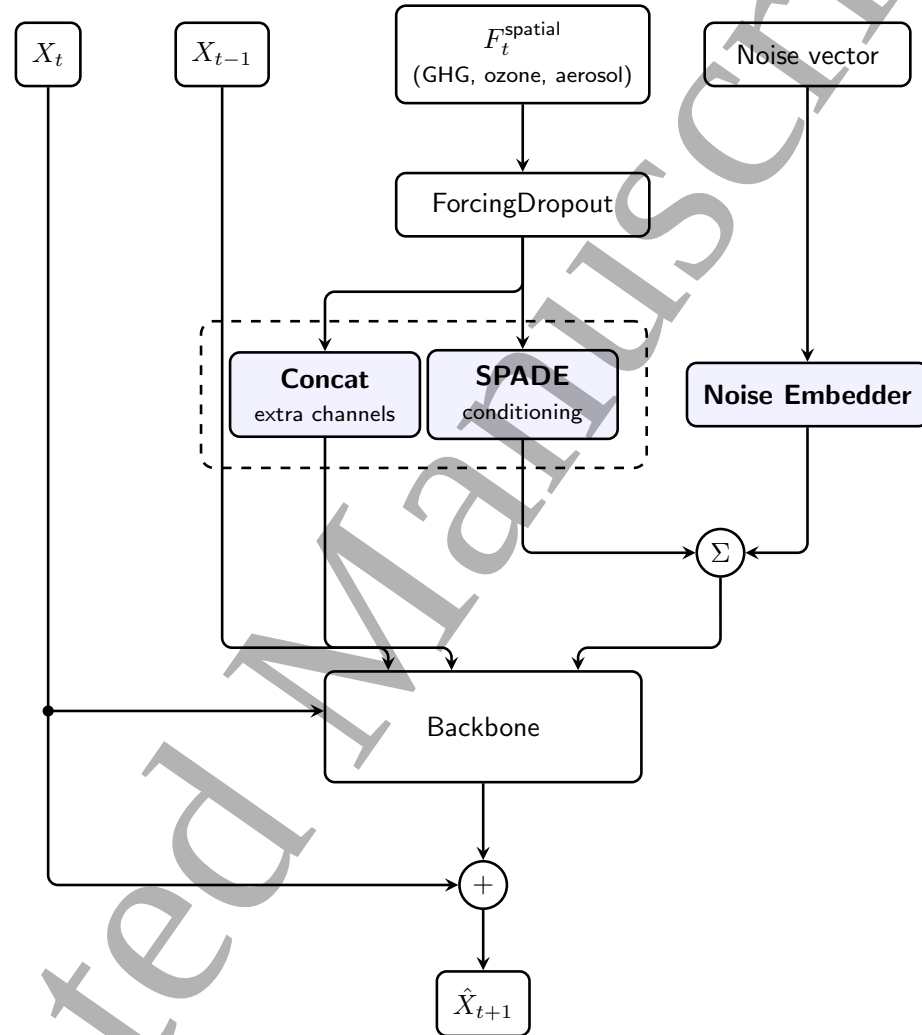


Figure 1: Architecture of the ArchesClimate model configured in a top-to-bottom pipeline. Top: Input preprocessing, **ForcingDropout**, and spatial conditioning modes. Middle: Aggregation of conditioning tokens (Σ). Bottom: ArchesClimate backbone receiving state inputs and conditioning.

scores each output channel independently and is therefore blind to whether the ensemble reproduces the correct joint structure across variables, levels, and grid points, the energy score is a genuinely multivariate generalization of CRPS and remains sensitive to that cross-dimensional dependence structure. We follow the same fair/unbiased finite-ensemble construction used for CRPS-trained ensemble weather models [28, 42, 69]. For a given input, the model produces M stochastic members $\{\hat{y}_1, \dots, \hat{y}_M\}$ by drawing M independent Gaussian noise vectors, each projected through a small learned embedder and added to the token-wise conditioning.

The energy score loss is then the batch-mean of the members' average distance to the target (the *skill* term) minus half their average pairwise distance to each other (the *spread* term):

$$\mathcal{L}_{ES} = \mathbb{E} \left[\frac{1}{M} \sum_{m=1}^M d(\hat{y}_m, y) - \frac{1}{2M(M-1)} \sum_{i=1}^M \sum_{j \neq i}^M d(\hat{y}_i, \hat{y}_j) \right] \quad (2)$$

The spread term is subtracted with an $M(M-1)$ (rather than M^2) normalization to exclude the zero self-distance terms, giving an unbiased estimator that does not shrink toward zero spread as M grows; at $M=1$ the spread term vanishes and $\mathcal{L}_{ES}$ reduces to the plain latitude-weighted RMSE. We use three members for our energy loss training.

3.4.2 Variogram Loss The energy score's spread term, as defined above, aggregates every variable/level/lat/lon-coordinate of the state into a single scalar distance per batch element. At the dimensionality of a full climate state, this aggregate concentrates tightly around its expectation regardless of how the underlying difference is spatially arranged. Spatially incoherent per-pixel noise and a spatially coherent, physically structured error pattern of the same total variance score almost identically [70]. To recover sensitivity to spatial structure, we add a Variogram Score term [71], evaluated on pairwise differences at a small, fixed set of spatial lags rather than on the full state at once. For a variable group v , spatial axis $\alpha \in \{h, w\}$, and lag ℓ , define the finite difference at power p :

$$\delta_{v,\alpha}^\ell(x) = |x_{v,\alpha+\ell} - x_{v,\alpha}|^p \quad (3)$$

The Variogram score is the mean, over variable groups, axes, and a lag set $\mathcal{L} = \{1, 2, 4\}$, of the fair (cross-member, bias-corrected) squared discrepancy between the predicted and ground-truth increments:

$$\mathcal{L}_{VS} = \text{mean}_{v,\alpha,\ell \in \mathcal{L}} \left[\frac{1}{M(M-1)} \sum_{i=1}^M \sum_{j \neq i}^M (\delta_{v,\alpha}^\ell(\hat{y}_i) - \delta_{v,\alpha}^\ell(y)) (\delta_{v,\alpha}^\ell(\hat{y}_j) - \delta_{v,\alpha}^\ell(y)) \right] \quad (4)$$

As with the energy score's spread term, the cross-member ($i \neq j$) form is an unbiased estimator of $(\mathbb{E}_m[\delta(\hat{y}_m)] - \delta(y))^2$ with no spread-shrinkage incentive; the naive single-member alternative $\text{mean}_m[(\delta(\hat{y}_m) - \delta(y))^2]$ would instead silently reward an artificially tight ensemble.

In practice, we use an explicit clamp to floor the Variogram score at 0 before it enters the loss. Like the energy score, $\mathcal{L}_{VS}$ is a fair estimator of the mean member-to-truth error minus half the mean member-to-member spread, and a fair/proper scoring rule allows this quantity to go negative, which happens whenever the ensemble's spread exceeds twice its skill. We clamp this at 0 rather than leaving it negative so the term can only ever be stopped from pushing further in an under-dispersed direction. The cost is that this gives up the estimator's strict properness: over-dispersion beyond the skill level is no longer penalized once past this floor, so the Variogram term alone provides no restoring force against excessive spread. In practice we accept this because the energy score's own spread term (section 3.4.1) still supplies that restoring force at the global scale; the Variogram term's job is specifically to add spatial-structure sensitivity, not to be the sole regulator of ensemble dispersion.

The total loss is $\mathcal{L} = \mathcal{L}_{ES} + w_{VS} \cdot \mathcal{L}_{VS}$, with w_{VS} ramped linearly from 0 to its target value (0.5) over the first 5,000 steps of Variogram training, so the model first learns to fit the broad state before being penalized on spatial structure.

To perform inference with AC-SSP, we generate climate state time series autoregressively, initialized by IPSL model conditions for a given SSP. For example, to emulate SSP4-3.4, the model is initialized using the first two months of the IPSL scenario (e.g., January and February 2015) and the corresponding forcings. AC-SSP then predicts the subsequent month's state, which is fed back into the model to extend the simulation. We define these generated time series as rollouts, with each rollout serving as a distinct ensemble member. A detailed visualization of this pipeline is available in [1, 13].

Hyperparameter	Value
Optimizer	Muon + AdamW (hybrid)
Muon Learning Rate	2×10^{-4}
AdamW Learning Rate	5×10^{-5}
Schedule	Cosine with Warmup (5,000 steps)
Warmup Steps	5,000
Betas (β_1, β_2), AdamW group	0.9, 0.98
Weight Decay	0.05

Table 2: Hyperparameters used in training the deterministic model. Values are reused from ArchesClimate and ArchesWeather. Optimization uses a hybrid scheme: Muon updates every ≥ 2 -dimensional weight matrix outside the embedder’s input/output boundary layers (the bulk of the backbone), while AdamW updates the embedder boundary layers (patch/de-patch projections, month/hour/timestep embedders, positional axes) and every low-dimensional (bias/norm) parameter, each at its own learning rate; weight decay applies to both groups except AdamW’s no-decay (bias/norm) parameters.

3.4.3 Compute and Running Time All training and inference reported in this paper run on a single NVIDIA H200 GPU (12 CPU cores, 250GB RAM). For the default $pf = 4$ deterministic model, training for 22,000 steps at batch size 4 took 6.5 h wall-clock time. The corresponding generative flow-matching model, also trained for 22,000 steps at batch size 4, took 5 h of wall-clock compute time. The energy-score model at $pf = 4$, trained to 22,000 steps at batch size 4, took approximately 6 hours, a comparable per-step rate to the non-energy-score models above, so the energy-score loss’s added cost at $pf = 4$ (three stochastic member forward passes during the pushforward window, section 4.1) is not the dominant factor in this run’s time. At inference, generating a full 85-year (1020 monthly steps) autoregressive rollout at our default settings took approximately 80 seconds for the deterministic model. The energy-score model’s rollout is the same as the deterministic model’s, at roughly 1.5 minutes for a single 85-year member.

3.5 Metrics

We use Latitude-Weighted Root Mean Squared Error (RMSE) and Normalized RMSE (for comparison across variables) to evaluate the accuracy of AC-SSP. To evaluate the variance of AC-SSP, we use Interannual Variability (IAV), Ensemble Spread and Spatial Standard Deviation. We use the Hydrostatic Balance as a physical sanity check when comparing ablated models. In meteorology, hydrostatic balance is expressed by combining the Hydrostatic Equation and the Ideal Gas Law [72]. As a second, independent physical sanity check, we use an energy-budget consistency check: we compute the column-integrated atmospheric energy (sensible, potential, kinetic, and latent contributions) and the upper-ocean heat content from the model’s own generated state, and check whether their combined rate of change over each rollout month is consistent with the model’s own net surface energy flux for that month, as expected from energy conservation [73, 74]. Because only the atmosphere and the upper ocean are represented, this is a consistency check and not a test of closure of the full Earth energy budget; we report the Mean Absolute Error of this residual as a global time series. Please see section A.8 for precise definitions of the listed metrics.

3.6 Ozone and Aerosol Analysis

The TCR-based reference used for the greenhouse gases (section A.10) assumes a linear relationship between forcing and warming, which cannot be established for ozone and aerosol. For these two groups of forcings, we run emulations of experiments that attempt to isolate both signals. For aerosol, we use *hist-piAer*, which holds aerosol forcings at pre-industrial levels while the rest of the forcings follow *historical* levels. For ozone, the only available ozone analysis for IPSL (excluding INCA-driven experiments) is *hist-stratO3* from the Detection and Attribution Model Intercomparison Project (DAMIP). This experiment holds all forcings at pre-industrial levels except for stratospheric ozone, which follows *historical* levels. We can therefore generate an emulation of the *historical* experiment (with all forcings included) and compare the difference of the generated *historical* experiment minus the forcing experiment for both the IPSL and AC-SSP. While it is not ideal that one experiment holds out the forcing and the other experiment holds out everything but the forcing, we are limited by the available data. Both experiments allow us to glimpse into how AC-SSP has learned to incorporate these forcings into long autoregressive

generations. We expect ozone held at pre-industrial levels to produce cooling effects and aerosol held at pre-industrial levels to produce warming effects, consistent with their respective roles in the historical forcing budget [54]. See section 4.3 for results of these experiments.

3.7 Baseline

We utilize MESMER-M [15, 16] as our primary baseline, a statistical climate model emulator designed to produce spatially explicit monthly temperature fields. While the original MESMER framework [75] focuses on annual temperature maps by accounting for both the long-term local warming signal and grid-point-specific internal variability, MESMER-M extends this capability to a monthly temporal resolution [15]. We selected MESMER-M as a benchmark because it is a well-established, non-machine-learning method that decomposes the climate signal into distinct mean and residual components.

The mean component of MESMER-M is modeled as a harmonic Fourier series that represents a smooth seasonal cycle. To account for non-stationarity, both the baseline and the seasonal amplitude are fitted as linear combinations of the annual mean temperature. The remaining residual component, which represents internal variability, is modeled using a first-order autoregressive process.

For our experimental setup, we calibrated MESMER-M using a single member from four different scenarios: SSP1-1.9, SSP1-2.6, SSP2-4.5 and SSP3-7.0, and the historical experiment. A single member suffices here because MESMER-M’s two fitted components – the harmonic seasonal-cycle/forced-response fit and the AR(1) internal-variability model – are each calibrated against the member’s own multi-decadal monthly time series, which already supplies hundreds of monthly samples per scenario; the AR(1) fit is designed to characterize the statistical properties of internal variability in general, not to memorize that one realization’s specific noise sequence, so it does not require multiple realizations to be sampled directly at calibration time. Anomalies were calculated relative to the 1850–1900 pre-industrial reference period. During the evaluation phase, we provided MESMER-M with the target annual mean surface temperatures from the IPSL model to predict the monthly temperature of two withheld scenarios, SSP4-3.4 and the overshoot scenario SSP5-3.4. Further technical details regarding the MESMER-M architecture can be found in [15].

4 Results

In this section, we evaluate the climate emulations produced by AC-SSP and show that it can accurately capture the spatial and temporal dynamics of different SSP scenarios. All reported results use the training hyperparameters that yielded the highest performance during ablation runs (see section 4.1). We compare to *MESMER* – *M* when possible. Following this statistical validation, we quantify the impact of each forcing on the generated climate states. We then assess the physical consistency of the generated states (subsubsection 4.2.1). After this, we analyze the SSP5-8.5 scenario and the abrupt-4xCO₂ scenario, which allows us to test the model’s capability for out-of-distribution generalization (section 4.4). Finally, we follow ClimateSet’s [39] reasoning that ML approaches should be tested across different climate models, and test the generalizability of AC-SSP on another climate model (CanESM5).

4.1 Architectural Ablations

We first fix a set of architectural and training-configuration choices by ablating them independently on the energy-score model. The push-forward length pf is varied while holding $p_{\text{all}} = 0.6$, and the forcing all-present probability p_{all} is varied while holding $pf = 4$ (see subsection 3.3); the $pf = 4$ and $p_{\text{all}} = 0.6$ rows therefore refer to the same run. We organize this into two ablation studies, each in its own subsection below: the forcing dropout parameters used during training and the push-forward rollout length.

4.1.1 Multi-step Finetuning We determine the optimal push-forward length (pf) by evaluating the energy-score model detailed in Table 3 (specifically comparing $pf \in \{2, 4, 8\}$, utilizing 5-member ensembles under the SSP4-3.4 scenario). As demonstrated in subsection A.3, this model exhibits sensitivity to initialization; we observe a 22 % RMSE spread across three otherwise identical training seeds. Consequently, performance differentials smaller than this baseline variance cannot be reliably distinguished from standard training stochasticity on the basis of single runs. The RMSE values for $pf = 2$ and $pf = 4$ yield a marginal difference of 0.0142 K, placing them within this threshold of uncertainty. However, they are distinctly separated by their inter-annual variability (IAV). The IAV for $pf = 4$ aligns closely with the IPSL reference target, whereas $pf = 2$

Ablation	Configuration	Temporally-Avg. RMSE (K)	Spatial Std. Deviation	Ensemble Spread	Inter-annual Variability
Pushforward Length pf	$pf = 2$	1.1831	13.5441	0.5717	0.9394
	$pf = 4$	1.1689	13.3068	0.3865	0.6961
	$pf = 8$	1.2726	13.6955	0.5558	1.0144
Forcing All-Present Probability p_{all}	$p_{all} = 0.0$	1.2278	13.3151	0.3868	0.8040
	$p_{all} = 0.2$	1.2226	13.3626	0.4819	0.8566
	$p_{all} = 0.4$	1.2310	13.4191	0.5561	0.8831
	$p_{all} = 0.6$	1.1689	13.3068	0.3865	0.6961
	$p_{all} = 0.8$	0.9941	13.3191	0.3640	0.6124
	$p_{all} = 1.0$	0.9800	13.3397	0.3476	0.6066

Table 3: Architectural and training-configuration ablations of the energy-score model, evaluated on the SSP4-3.4 validation scenario. Lower is better for RMSE; IPSL’s own spatial standard deviation and inter-annual variability over the same final-decade window are 13.2520 and 0.6337 respectively (Table 4), for reference. Each row is a separate training run; the push-forward sweep uses $p_{all} = 0.6$ and the p_{all} sweep uses $pf = 4$, so the $pf = 4$ and $p_{all} = 0.6$ rows are the same run. All runs use $p_{drop} = 0.2$. The RMSE column is temporally averaged: monthly tas is first averaged to annual means, then RMSE (K) is computed against the IPSL target pooled (lat-weighted) over the last 10 years. All models are 5-member ensembles.

overestimates it. This discrepancy (0.24) is large compared with the seed-to-seed spread of IAV (≈ 0.02 , subsection A.3), favoring the $pf = 4$ configuration.

Conversely, $pf = 8$ records the highest RMSE and IAV, alongside a spatial standard deviation that diverges furthest from the IPSL baseline. Its RMSE differs from that of $pf = 4$ by 0.104 K, which is smaller than the seed-to-seed spread, so this ranking is suggestive rather than conclusive; on the basis of these results we do not pursue $pf = 8$. We therefore establish $pf = 4$ as our default configuration for all subsequent analyses.

4.1.2 Forcing Dropout Parameters This ablation investigates the fundamental trade-off governed by the all-present probability (p_{all}): insufficient exposure to the complete forcing set prevents the model from integrating all forcings jointly, whereas inadequate exposure to partial or missing forcings precludes robust evaluation when forcings are deliberately withheld.

Evaluating six p_{all} configurations with $pf = 4$ (and $p_{drop} = 0.2$) using 5-member ensembles (Table 3) reveals a broad improvement in root-mean-square error (RMSE) as p_{all} increases. Configurations where $p_{all} \in \{0.0, 0.2, 0.4, 0.6\}$ exhibit an RMSE of approximately 1.17-1.23 K, while the $p_{all} \in \{0.8, 1.0\}$ configurations achieve lower RMSE, with RMSE of 0.99 K ($p_{all} = 0.8$) and 0.98 K ($p_{all} = 1.0$). Within this top-performing pair, however, the margin is negligible. As established in subsection A.3, this gap falls well within the 22 % (approximately 0.22 K) seed-to-seed variance for this model family, rendering the two configurations indistinguishable on the basis of a single training run’s RMSE. The gap between these two configurations and the remaining four (≈ 0.2 K) is of the same order as the seed-to-seed spread, so the RMSE ranking should be read as indicative only.

To resolve this ambiguity, we evaluate $p_{all} = 0.8$ and $p_{all} = 1.0$ against physical criteria, employing single-forcing piControl-held attribution ablations referenced against the ERF/TCR estimates (see section A.4 for a comprehensive per-gas decomposition, and see section A.10 for a definition of the ERF/TCR estimates). Three distinct metrics favor the $p_{all} = 0.8$ configuration. First, it correctly captures the sign of the methane response, whereas the $p_{all} = 1.0$ configuration incorrectly simulates a warming response (+0.332 K). Second, both configurations end the century with an ozone response of the incorrect sign (warming of about +1.3 K at $p_{all} = 0.8$ and +1.1 K at $p_{all} = 1.0$ by 2100), but the trajectories differ: at $p_{all} = 0.8$ the response has the expected sign (cooling of up to about 1.5 K) until around 2050 and only then reverses, whereas at $p_{all} = 1.0$ it is warming from shortly after the start of the rollout. Finally, while both exhibit significant N₂O over-attribution, the magnitude is smaller for $p_{all} = 0.8$ ($\times 9.27$ the TCR reference versus $\times 12.41$). The sole exception is CO₂, where $p_{all} = 0.8$ deviates slightly further from the expected response (0.129 K versus 0.063 K); however, this represents the smallest deviation among the four gases and does not negate the advantages observed elsewhere. Consequently, the physical forcing-attribution evidence supports $p_{all} = 0.8$, despite the aggregate RMSE parity. We thus select $p_{all} = 0.8$ as the default checkpoint for AC-SSP and all subsequent analyses in this paper.

To confirm the necessity of the forcing signal, we evaluated a baseline model trained entirely without forcing conditioning. While the autoregressive generation remained numerically stable and

captured a generic warming tendency, the state trajectory diverged significantly from the true climate path. This suggests that external forcing information is needed: in this experiment the model could not adequately reconstruct the forced response from the unconditioned state dynamics alone.

4.2 Comparison to MESMER-M baseline

Our model demonstrates a high capacity to accurately emulate the IPSL outputs for several SSP experiments. In Table 4, we compare *AC-SSP* against the baseline *MESMER-M* across several metrics applied to global land surface temperature for the decade 2090-2100. *AC-SSP* is a 5-member ensemble, as is *MESMER-M*; the IPSL reference data provides only a single member for the SSP4-3.4 and SSP5-3.4 scenarios. Performance is evaluated using RMSE for accuracy, spatial standard deviation to quantify geographic temperature gradients (e.g., Arctic vs. Saharan contrasts), Inter-annual Variability (IAV), and ensemble spread to compare internal variability. As there is only a single ground-truth member, we need to be cautious of overstating the performance of RMSE, given the unknown extent of internal variability in the target. All metrics are derived from yearly averages spanning 85 years for SSP4-3.4 and 60 years for SSP5-3.4.

AC-SSP's RMSE is comparable to *MESMER-M*'s on both scenarios. On SSP5-3.4, *AC-SSP* is in fact the lower-error of the two. Spatial standard deviation also tracks the IPSL target closely for both. *AC-SSP*'s ensemble spread is consistently larger than *MESMER-M*'s, consistent with *AC-SSP*'s autoregressive rollout accumulating more member-to-member divergence over 85 years than *MESMER-M*'s statistical emulation does. Interestingly, *AC-SSP*'s IAV is closer to the IPSL target's than *MESMER-M*'s is, in both scenarios. Because *MESMER-M* is conditioned on prescribed global temperatures, its inter-annual variations might be expected to align closely with the IPSL target by construction; that *AC-SSP*'s IAV tracks the target at least as well suggests the autoregressive rollout is not simply smoothing away real interannual variability, despite having only a short (two-month) input context to draw on. Note that because IPSL provides only a single member, no reference ensemble spread is available for comparison.

Scenario	Model	RMSE (K)	Spatial Std. Deviation	Ensemble Spread	Inter-annual Variability
SSP4-3.4	<i>AC-SSP</i>	0.9843	13.3329	0.3971	0.6161
	<i>MESMER-M</i>	0.9390	13.2717	0.1179	0.7685
	IPSL	n/a	13.2520	n/a	0.6337
SSP5-3.4	<i>AC-SSP</i>	0.9739	13.4434	0.3928	0.6050
	<i>MESMER-M</i>	1.0963	13.6192	0.0999	0.6375
	IPSL	n/a	13.6069	n/a	0.5166

Table 4: Performance of land surface temperature across metrics and SSP scenarios, for *AC-SSP* against *MESMER-M* and the IPSL target. All values are based on annual averages of land surface temperature for the decade 2090-2100, as *MESMER-M* is a land-surface temperature model only. See subsection A.8 for definitions of the metrics. Ensemble Spread is unavailable for IPSL as it has only one ensemble member for SSP4-3.4 and SSP5-3.4.

4.2.1 Physical Consistency An autoregressive emulator trained on CMIP output is not constrained to obey the conservation laws of the system it emulates. We use two diagnostics to probe this for the ArchesClimate emulator: hydrostatic balance (the relationship between air temperature (ta) and geopotential height (zg)) and a coupled atmosphere-ocean energy budget. See section A.8 for definitions of both diagnostics.

Results here are for a 5-member ensemble of *AC-SSP*, evaluated with the SSP 5-3.4 test scenario. As illustrated in Figure 2, a residual near zero indicates adherence to hydrostatic equilibrium. Rather than treating any nonzero residual as an outright violation of physical law, we treat this diagnostic as a measure of consistency with the climate model it is trained against.

The vertical profile in Figure 2 shows that the hydrostatic residual is small relative to the thickness of the layers it is computed over (mean absolute residual 14.3 m for *AC-SSP* and 6.2 m for the IPSL target, against layer thicknesses of roughly 0.7–4.5 km), indicating a robust modeled relationship between temperature (ta) and geopotential (zg). A localized inconsistency appears near 40 hPa in the Tropics, plausibly corresponding to sharp thermal gradients, though the

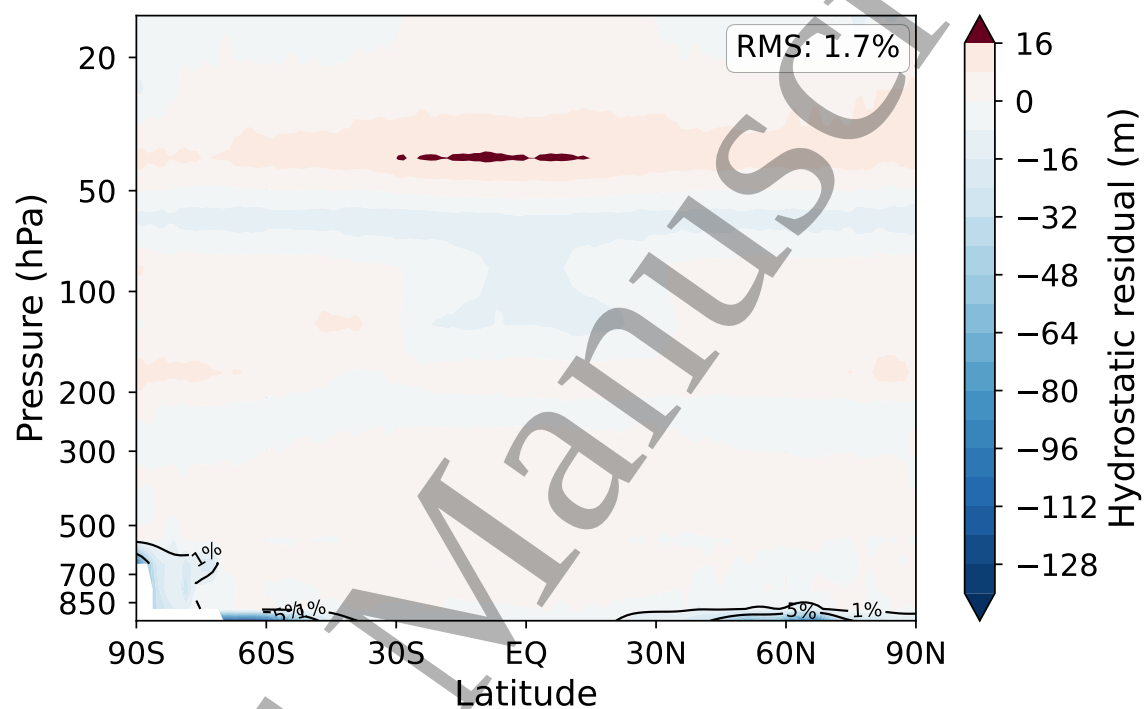


Figure 2: Hydrostatic balance residual, zonal and 60-year-rollout mean (computed locally per grid cell, then averaged), for *AC-SSP*, *SSP5-3.4*. Values closer to zero are better. The y-axis is pressure (surface at the bottom), log-scaled; the x-axis is latitude in degrees, with faint dotted reference lines at the standard 90S-90N latitudes. Black contour lines (lightly smoothed along latitude for legibility) mark where the residual reaches 1%/5% of the local hydrostatic layer thickness, and the boxed annotation reports the area-weighted root-mean-square (RMS) of that percentage across the whole column. The hydrostatic residual is small in both absolute and relative terms (area-weighted RMS 1.7% of layer thickness) and confined almost entirely to the surface and a narrow tropical band near 40 hPa, indicating a largely physically consistent temperature-geopotential relationship elsewhere in the column.

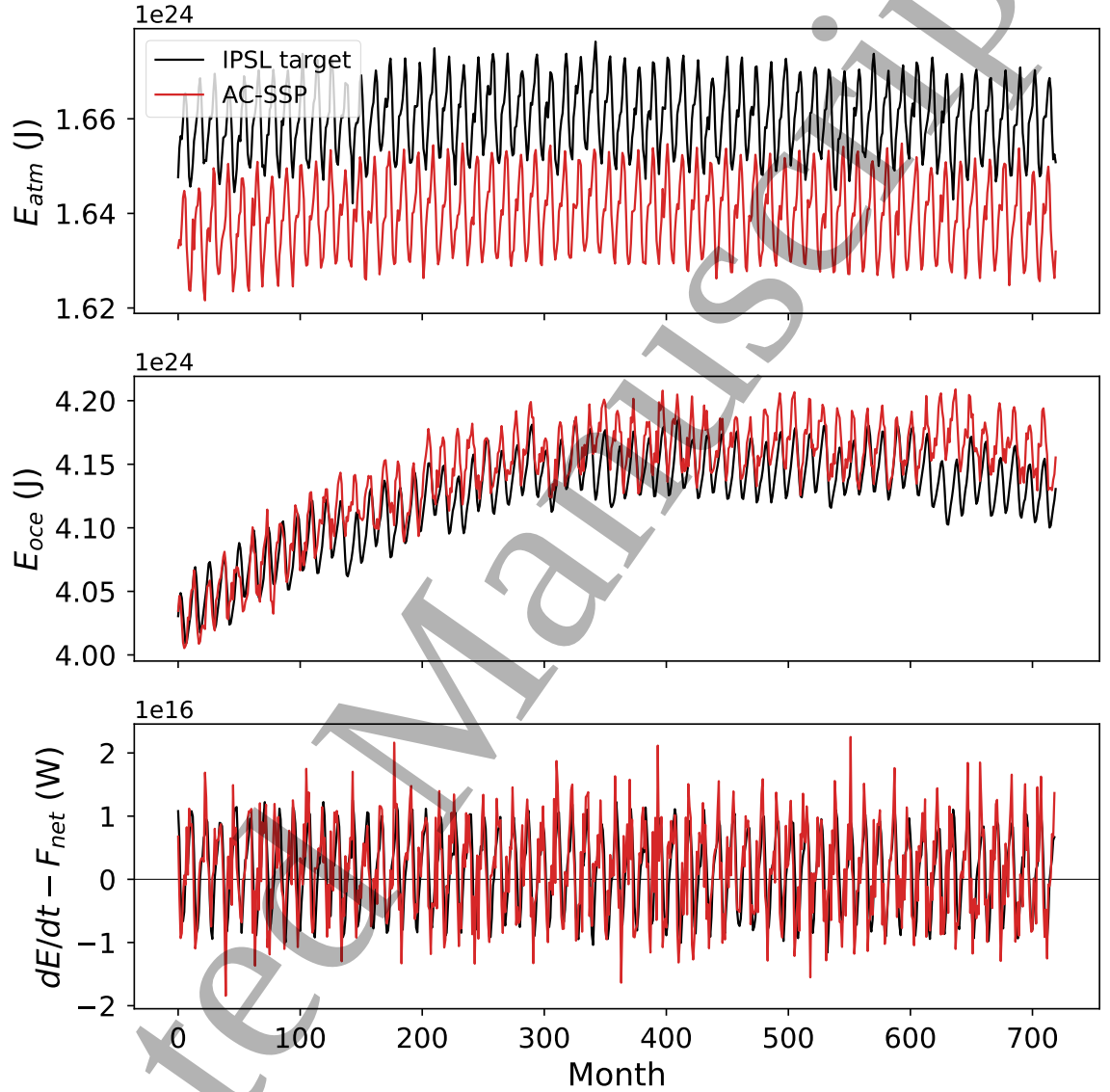


Figure 3: Coupled atmosphere–ocean energy budget for *AC-SSP*. Energy budgets for SSP5-3.4: global E_{atm} and E_{oce} trajectories, model vs. target, and the $dE/dt - F_{net}$ energy-budget residual. E_{atm} is the column-integrated atmospheric energy, the sum of sensible (internal), potential, kinetic, and latent heat contributions (eq. (17)); E_{oce} is the upper-ocean heat content, the depth-integrated seawater temperature down to 147.4 m (eq. (19)). E_{atm} and E_{oce} both closely track their targets with a consistent offset rather than diverging.

absolute error remains low. More substantial residuals emerge at the surface, plausibly driven by a combination of unresolved geopotential gradients at the land-atmosphere boundary; attempts to mitigate this via logarithmic transformations of zg yielded no measurable improvement in physical consistency (not shown). In the atmosphere–ocean energy budget (Figure 3), the atmospheric energy (E_{atm}) closely tracks the target but exhibits a systematic offset. The total E_{atm} mean absolute error is $2.74 \times 10^7 \text{ J/m}^2$. Computing the mean absolute error of each of its four terms separately gives shares of this total of $\approx 58\%$ for sensible heat, $\approx 33\%$ for latent heat, $\approx 16\%$ for potential energy and $\approx 1.5\%$ for kinetic energy. These shares sum to about 108% because the signed errors of the individual terms partly offset each other in the total, so the term-wise errors are not additive. Please see section A.9 for the calculations of these numbers. This error distribution is consistent with a vertically non-uniform temperature bias: the model produces a tropospheric warm bias (+0.2 to +0.6 K) that reverses into a stratospheric cold bias (−0.2 to −0.8 K). Because E_{atm} is a mass-weighted column integral, the denser troposphere dictates the net positive error. Both diagnostics suggest that AC-SSP’s departures from physical consistency are small and systematic, and plausibly tied to a single identifiable source (a vertically-structured temperature bias). The energy-budget residual points in the same direction: its mean is $1.47 \times 10^{15} \text{ W}$ for AC-SSP and $1.26 \times 10^{15} \text{ W}$ for the IPSL target evaluated identically, so most of the residual is a property of the incomplete budget and not of the emulator.

Finally, the ocean energy trajectory (E_{oce}) exhibits slight drift and a model-target mean absolute error of $3.6 \times 10^8 \text{ J/m}^2$. This drift is plausibly related to the restricted depth of the emulated ocean (147.4 m is the deepest level represented; subsection A.1), which lacks the deep-ocean reservoir necessary to facilitate transient heat uptake [76].

4.3 Forcing Response: Physical Consistency of Single-Forcing Removal

We next check whether AC-SSP attributes the correct sign and magnitude of warming to each forcing, at the surface and across the atmospheric vertical profile. For each forcing (see section 3.1 for a complete description), we hold that channel at its piControl (pre-industrial) value for the entire SSP5-3.4 rollout while leaving every other forcing untouched, and compare the resulting yearly global-mean *tas* trajectory against the model’s own all-forcings baseline. We hold forcings at piControl levels specifically to be consistent with DAMIP experiment design. We use the TCR procedure described in section A.10 for CO_2 , N_2O , and CH_4 . For ozone and aerosols, we evaluate with explicit experiments, *hist-piAer* and *hist-stratO3*. Details for this procedure can be found in section 3.6.

Figure 4 illustrates the forced response learned by AC-SSP. Holding CO_2 at piControl levels cools the trajectory by 2.00 K by 2100. This exceeds the 1.5 K cooling predicted by the TCR reference, indicating that the model over-attributes aggregate warming to CO_2 specifically. The methane response, conversely, closely tracks the estimated TCR. The most severe discrepancy occurs with N_2O : withholding N_2O induces a 2.0 K cooling by 2100, overshooting the 0.2 K TCR estimate by a factor of ten.

We hypothesize that this over-attribution arises from strong collinearity between the CO_2 and N_2O forcing trajectories within the training corpus. Because CO_2 and N_2O concentrations rise synchronously across the SSP scenarios, the historical and SSP simulations primarily constrain the aggregate greenhouse gas response rather than isolating individual constituent contributions. The idealized *1pctCO2* experiment would in principle anchor the isolated CO_2 signal, but it is not part of our training set, and no equivalent experiment exists for N_2O . The model may therefore fail to establish an independent physical sensitivity for N_2O , defaulting instead to a confounded correlation. Resolving this degeneracy and establishing a robust N_2O representation would likely require expanding the training dataset to include explicit single-forcing N_2O ablations.

The same piControl-held ablations, resolved across the atmospheric column rather than collapsed to a global-mean surface time series, are shown in Figure 5, for AC-SSP for SSP5-3.4. Removing CO_2 or N_2O produces the largest and most vertically coherent response of the five forcings, and both share the same characteristic shape: tropospheric cooling (roughly −1.2 to −2.0 K at 1000 – 500 hPa) that reverses sign through the upper atmosphere into a pronounced stratospheric warming (+0.7 to +0.8 K at 50 hPa, rising to +1.6 to +4.2 K at 10 hPa). This sign reversal is physically expected: removing a well-mixed greenhouse gas removes both its tropospheric warming effect and its stratospheric cooling effect, so the piControl-held column develops a warmer stratosphere alongside a cooler troposphere. CO_2 ’s effect is somewhat larger than N_2O ’s at most levels and regions, consistent with its larger radiative forcing contribution over this period. Methane, by contrast, produces a negligible response at every level and region, despite showing a more substantial surface response in Figure 4.

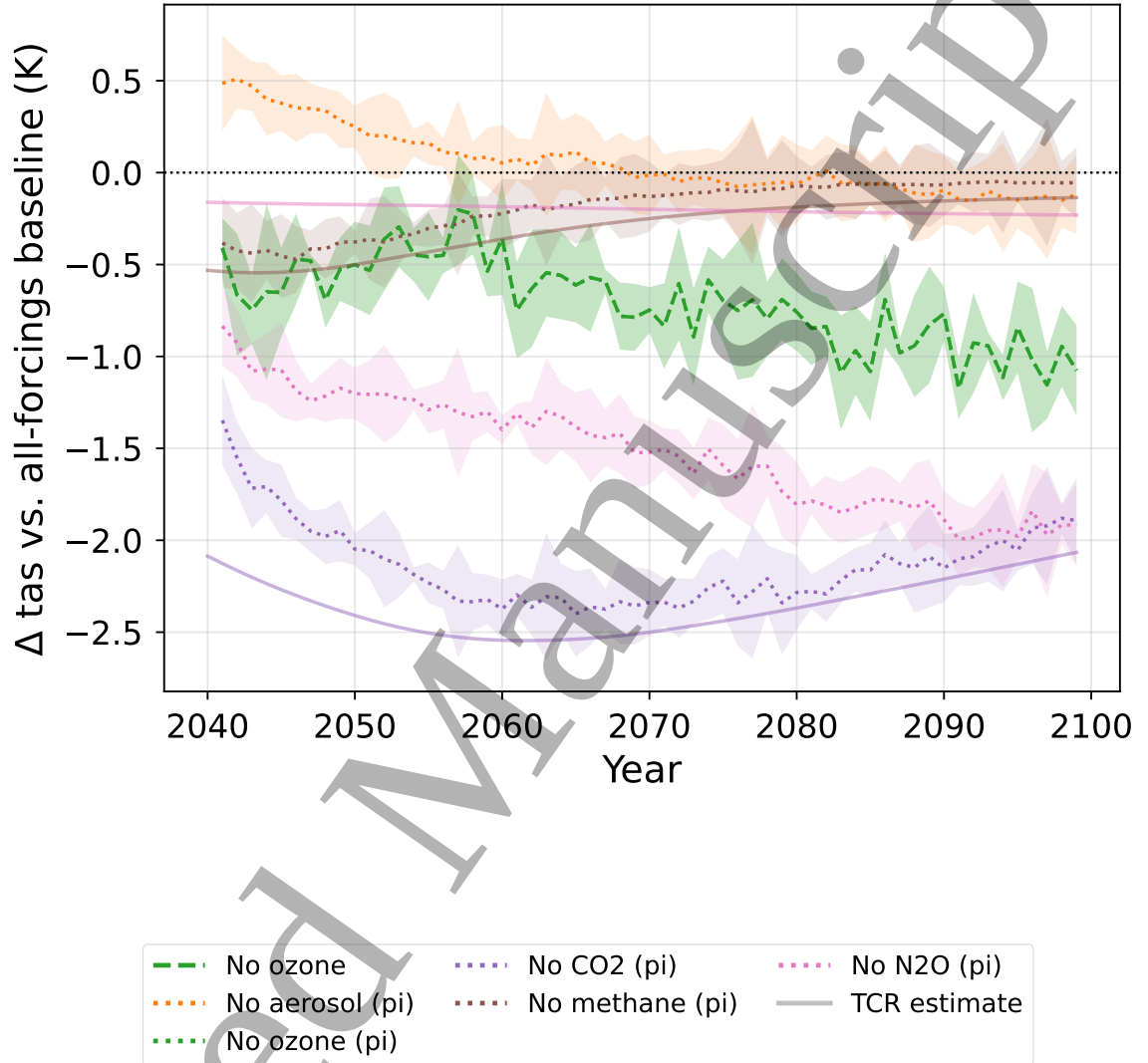


Figure 4: Single-forcing piControl-held ablations under SSP5-3.4-over, *AC-SSP*: yearly global-mean *tas* difference from the all-forcings baseline (dotted; mean across members, shaded band is ± 1 std across members). Each forcing is set to pre-industrial levels and held, the rest are left the same. For CO_2 /methane/ N_2O , the same-colored solid, semi-transparent line is that gas's physically expected magnitude from the ERF formula scaled by the TCR-derived sensitivity λ_{TCR} ; no aerosol/ozone reference is shown, as the simple ERF formulas used here do not cover those forcings. Note the shaded band here is a single-year member spread and so includes each member's own interannual variability; it is not directly comparable in magnitude to Figure 5's decade-averaged band. *AC-SSP* recovers the correct sign of every forcing's contribution, but over-attributes warming to CO_2 and especially N_2O relative to their TCR references, while methane's response matches closely.

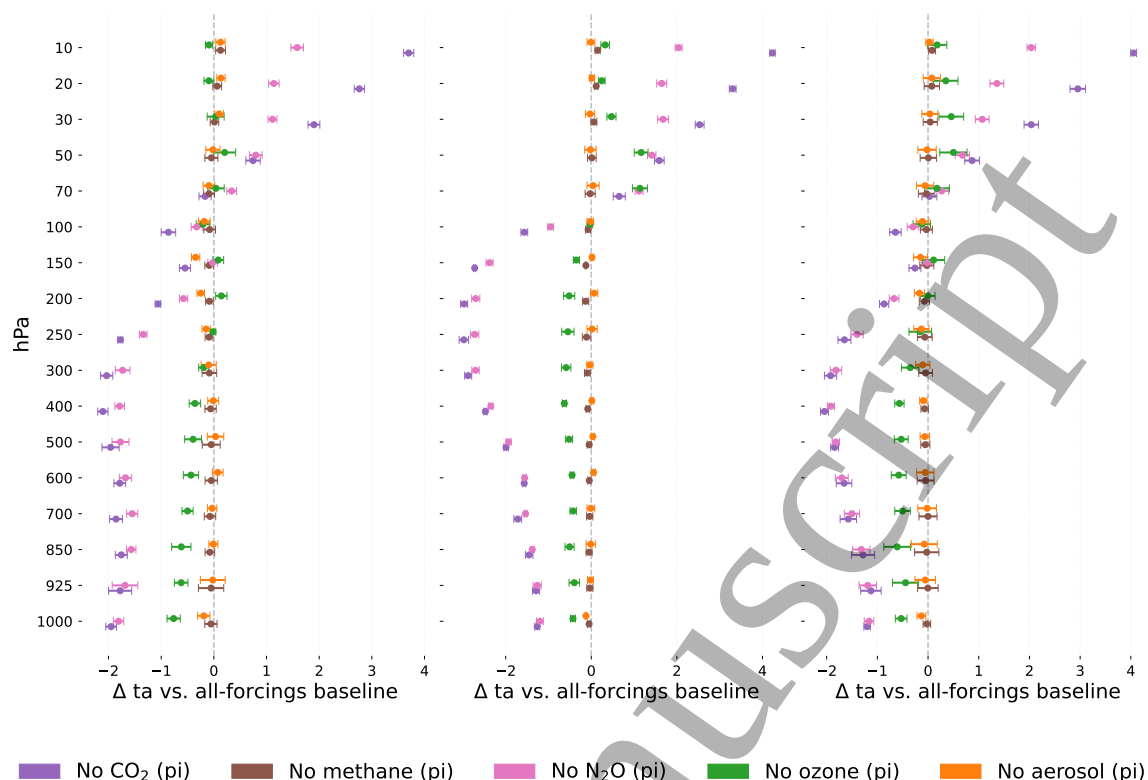


Figure 5: Atmospheric-column response to each single-forcing piControl-held ablation, *AC-SSP*, SSP5-3.4 averaged over the last decade (2090-2100): area-weighted mean ta difference from the model's own all-forcings baseline, by pressure level, left to right North (18.3°N-90°N) / Tropics (19.5°S-18.3°N) / South (90°S-19.5°S) latitude bands. Shaded bands show ± 1 std across ensemble members, computed after averaging each member over its own trailing decade. This suppresses independent year-to-year noise relative to Figure 4's single-year band. Removing CO₂ or N₂O produces the same physically-expected sign reversal (tropospheric cooling, stratospheric warming) at every latitude band, while methane and aerosol produce comparatively little vertical-column response.

Ozone exhibits a modest tropospheric cooling effect (-0.4 to -0.8 K at 1000 hPa). Ozone changes involve several interacting processes with divergent climate impacts, which the single-forcing ablation used here cannot separate. Regarding aerosols, our current representation is restricted to a two-dimensional atmospheric load rather than a vertically explicit column. However, because the aerosol forcing signal under the SSP5-3.4 scenario remains minimal for the 2090-2100 period, the deviation from pre-industrial control (piControl) levels is expected to be negligible.

To provide analysis of the forced response of ozone and aerosols in *AC-SSP*, we test against two physically consistent single-forcing trajectories, rather than a piControl-held ablation of an all-forcings baseline as performed above. A complete description of this experiment is provided in Section 3.6. As illustrated in Figure 6, the aerosol forcing experiment generated by *AC-SSP* closely tracks the smoothed mean trend of the IPSL and scales appropriately throughout the historical period. Conversely, the modeled ozone response captures the trend effectively until the latter portion of the simulation, where the IPSL signal becomes stronger than the *AC-SSP* signal. This divergence may stem from an incomplete representation of stratospheric ozone, as our configuration incorporates only a subset of available forcings (10 out of the 66 layers, spread across the atmospheric column, are included as forcings). Nevertheless, capturing the positive trend suggests that the influence of stratospheric ozone on the surface is at least partly represented. Precipitation exhibits a comparable trajectory, with similar deterioration at the end of the stratospheric ozone experiment.

Interestingly, internal variability is highly pronounced from the onset of each IPSL experiment, presumably reflecting an initial spin-up period. Our emulator requires substantial time to properly generate sufficient internal variability between the two runs, as they are initialized with the same noise token and only differ in the forcing held to pre-industrial levels. While both states initialize with marginal discrepancies, these differences amplify over time, becoming particularly pronounced in the precipitation fields. Implementing a dedicated spin-up period for the emulator could address

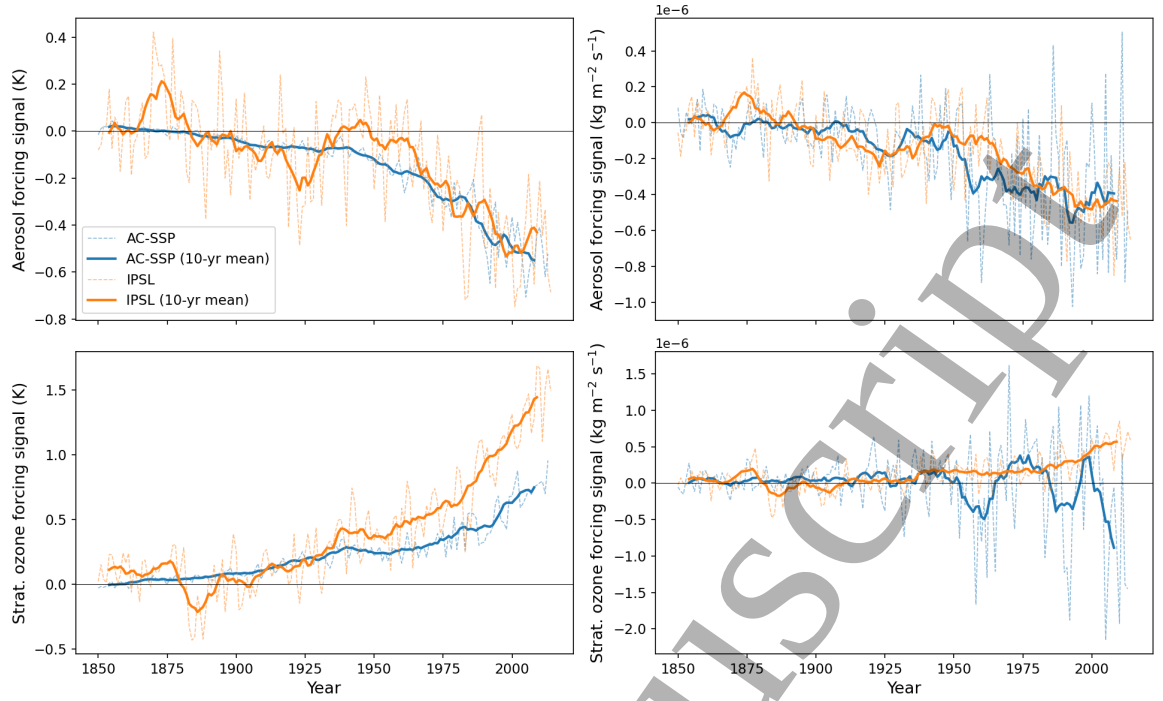


Figure 6: *AC-SSP* vs. *IPSL*, 1850-2014: (top row) aerosol forcing signal, generated *historical* minus generated *hist-piAer* global-mean value, against the same difference computed from the real *IPSL* data; (bottom row) stratospheric-ozone forcing signal, analogous but for *hist-stratO3* (an additive DAMIP construction, so it reflects the combined effect of every non-strat-ozone forcing, not stratospheric ozone alone – see subsection 4.3). Left column: *tas* (K); right column: *pr* (kg m⁻² s⁻¹). Because both the real, single-realization *IPSL* series and *AC-SSP*’s own generated series are noisy year to year at this global-mean resolution, each panel shows a 10-year centered rolling mean of each series (solid) alongside the corresponding raw one (faint dashed). Despite never training on either DAMIP experiment, *AC-SSP*’s smoothed trend reproduces the sign and long-term trend of both real *tas* forcing signals, tracking the smoothed *IPSL* trend closely for aerosol and underestimating the ozone signal toward the end of the period; the *pr* signal is far noisier for both series and *AC-SSP*’s precipitation response is less cleanly separated from noise than its temperature response.

this early variance. Alternatively, a more engineered solution would involve applying a distinct noise-conditioning token between the two experiments.

4.4 Out-of-Distribution Validation

This section evaluates out-of-distribution generalization from two angles: SSP5-8.5, a smooth shift in CO₂ forcing beyond anything seen in training, and the idealized abrupt-4xCO₂ experiment, a discontinuous forcing step. The first one probes whether the model extrapolates along a familiar direction pushed further than it has seen, the other probes whether it can respond correctly to a forcing discontinuity unlike anything in its training trajectories.

As the most extreme high-emission pathway in the CMIP6 suite, SSP5-8.5 represents a climatic regime characterized by high CO₂ concentrations that provide a high climate change signal-to-noise ratio [77]. Here, we use this experiment to assess how specific forcing agents influence model extrapolation when greenhouse gas forcing conditions exceed value ranges seen during training. The following analysis focuses on density plots generated for the 2090–2100 period, as well as rollouts for the entire period.

The density plots for the stable base models, shown in the lower panel of Figure 7, show a multimodal distribution of surface temperature at each grid cell. We verified each mode’s identity directly, by checking which grid cells (and their latitudes) actually populate each mode’s temperature window. The four modes are: the Antarctic interior (coldest, elevation-driven), Antarctic’s coastal/peripheral regions (second-coldest), the Arctic (warmer than either Antarctic mode, since it is ocean/ice-moderated with no comparable elevation effect), and the tropics (warmest); mid-latitudes populate the smooth area between the Arctic and Tropic modes. Of the four modes, the Arctic mode is the least accurate, and we attribute this to a combination of factors

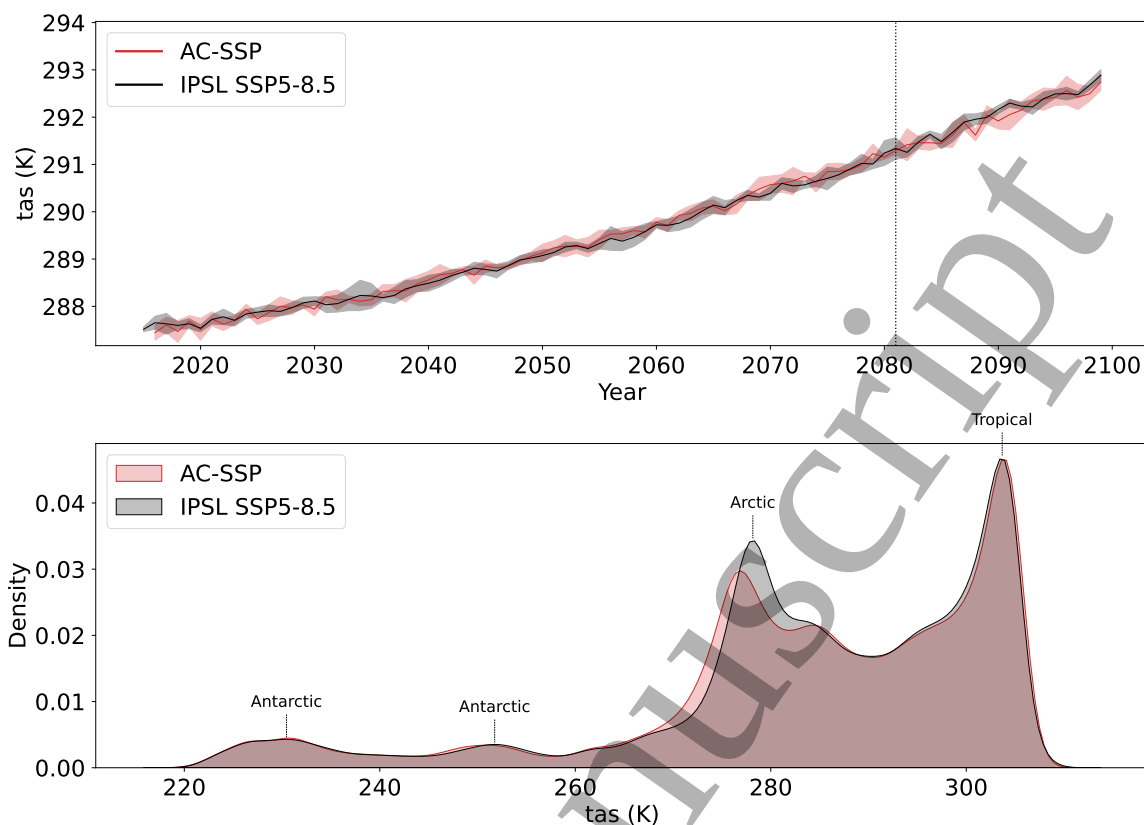


Figure 7: Out-of-distribution performance under the SSP5-8.5 high-emission scenario, *AC-SSP* vs. the IPSL target, a 4-member model ensemble vs. the 4-member IPSL ensemble. (Top) Time series of global-mean yearly *tas* (lat-weighted spatial average over all grid cells, then a mean $\pm$ std band across members/realizations) from 2015-2100; the dotted vertical line marks 2081, the first year SSP5-8.5's CO_2 concentration exceeds the maximum CO_2 level reached anywhere in the training data (SSP3-7.0's own maximum, reached at its final year, 2100). (Bottom) Probability density functions (PDFs) of grid-cell yearly-mean *tas*: computed by pooling every grid cell, member/realization, and year of the final decade into a single distribution, with each of the four modes labeled by the temperature regime it corresponds to (Antarctic interior, Antarctic coastal/peripheral, Arctic, Tropical). *AC-SSP* tracks the IPSL target's warming trend and grid-cell temperature distribution closely across the whole 85-year rollout and all four modes, with the Arctic mode's density being the only shortfall.

rather than a single cause. First, the extreme forcing of the SSP5-8.5 pathway pushes the Arctic furthest outside the model's training distribution. Because the region warms disproportionately faster than the global mean due to Arctic amplification [78], end-of-century grid-cell *tas* values reach magnitudes entirely absent from the historical and moderate SSP training sets (subsection A.1). Second, the polar regions also carry substantially higher year-to-year variability than the tropics or mid-latitudes, because of sea-ice/open-ocean transitions. Third, this effect is further increased due to the fact that a given grid cell spans less physical distance at high latitudes, so the same fixed-resolution grid resolves comparatively finer, noisier spatial structure there; a model trained with a loss that is latitude-aware will tend to under-weight this high-variance, small-area region relative to the tropics, where each grid cell covers far more physical area and dominates a weighted loss. We currently apply latitude-area weighting in the training loss, which we expect compounds the Arctic mode's underrepresentation.

4.4.1 Abrupt-4xCO₂ As a complementary stress test, we evaluate the models using the idealized abrupt-4xCO₂ experiment. Here, atmospheric CO₂ is instantaneously quadrupled from its 1850 pre-industrial baseline and held constant, confronting the model with a forcing discontinuity rather than the continuous, physical ramps present in the training data (subsection A.1). We assess the global-mean near-surface air temperature (*tas*) trajectory against the IPSL target, holding all non-CO₂ forcings at 1850 levels. To determine whether the energy-score objective itself degrades the transient response, characterized by rapid initial warming followed by a slow, ocean-mediated equilibration, we compare *AC-SSP* against a deterministic baseline ($pf = 4$, MSE loss, single

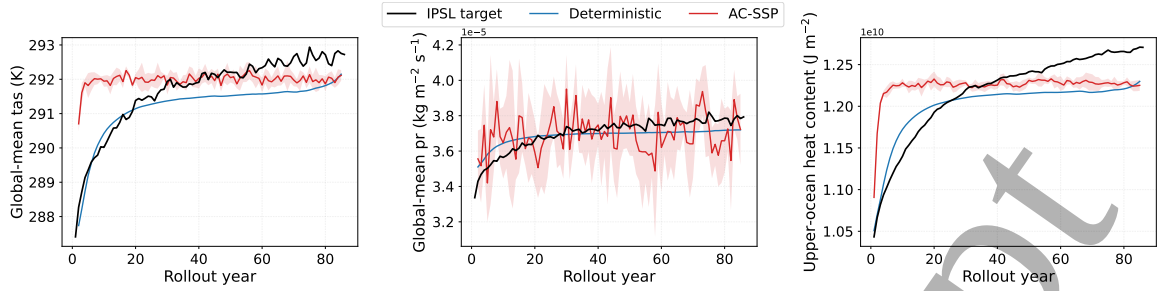


Figure 8: Abrupt-4xCO₂, global-mean *tas* (left), *pr* (middle), and upper-ocean heat content (right, depth-integrated as in eq. (19)) over an 80-year rollout, against the IPSL target (r2ilp1f1, black): the deterministic model (with all other parameters held constant) versus AC-SSP. The deterministic model reproduces the target’s slow, ocean-mediated transient warming in all three quantities, while AC-SSP instead jumps almost immediately to a near-final state and plateaus.

ensemble member).

The deterministic baseline captures the slow signal: global-mean *tas* rises smoothly by 3.50 K (from a mean of 288.28 K during years 1-5; to a mean of 291.78 K during years 76-80), mirroring the target’s trajectory despite underestimating the final magnitude. AC-SSP, conversely, effectively bypasses the transient phase entirely. It reaches within 1 K of its final state in the first five years and plateaus, warming by only 0.98 K over the remainder of the period. Global-mean precipitation (*pr*) exhibits an identical failure mode, with AC-SSP capturing only half of the deterministic model’s smooth rise alongside exacerbated inter-annual noise (Figure 8, middle panel). Upper-ocean heat content (Figure 8, right panel; computed as in section A.9) shows the same phenomenon. This indicates that adding ocean variables to an energy-score-trained model does not by itself let it learn slower-than-atmosphere temporal dynamics, whereas the deterministic model with the same variables does.

AC-SSP does not asymptotically approach an equilibrium; it immediately jumps to its final state. Because both models utilize identical architectures, this discrepancy most plausibly stems from the probabilistic objective. We hypothesize that the energy score, by incentivizing a state distribution conditioned strictly on the concurrent forcing, biases the model toward a forcing-conditioned equilibrium rather than a time-lagged trajectory. The deterministic model, unconstrained by distributional requirements, preserves the lagged trajectory learned during push-forward finetuning.

If this is correct, fixing this requires addressing the energy-score loss. We propose three untested but physically motivated directions: modeling the energy score as a residual atop a frozen deterministic base trajectory, scoring the push-forward rollout over the full temporal trajectory rather than frame-by-frame to penalize implausible paths (i.e., paths that quickly jump to some equilibrium) or conditioning the model on an extended historical state window to provide explicit rate-of-change information.

4.5 CanESM5 application

To assess whether AC-SSP generalizes beyond the IPSL-CM6A-LR dataset, we adapt the preprocessing pipeline for CanESM5 [79], an alternative CMIP6 model. This model was selected because it had the dataset most similar to IPSL among available CMIP models. This adaptation requires several structural changes prior to model training. The level grid, the ocean vertical grid, the surface variable set, and the available experimental protocols are all adjusted to fit the CanESM5. Please refer to section A.11 for details on the data and preprocessing of the CanESM5 model.

Figure 9 illustrates the single-forcing attribution test applied to the CanESM5 energy-score model under the SSP5-3.4 scenario. The results indicate that despite architectural differences in the base model, the learned dynamics remain consistent with the IPSL baseline: holding other forcings constant results in relative cooling, and the CO₂, CH₄, and N₂O responses each track their own ERF×TCR reference (using CanESM5’s own 2.8K TCR, section A.10) with the same qualitative over-attribution to CO₂/N₂O and close methane agreement seen for IPSL. Ozone has no comparable quantitative reference (section A.10), so that trace remains only qualitatively comparable to the IPSL results. Holding ozone at pre-industrial levels yields cooling trajectories similar to those observed in the IPSL model. Ultimately, these observations suggest that the

framework captures comparable underlying dynamics even when applied to an ESM with differing physical parameterizations.

Figure 10 shows the performance of the CanESM5 energy-score model across three distinct evaluation scenarios: SSP4-3.4 (in-distribution validation), SSP5-3.4-over (an out-of-distribution overshoot pathway absent from the training set), and SSP5-8.5 (the extrapolation stress test previously applied to the IPSL model in subsection 4.4). Consistent with the IPSL results, the CanESM5 model tracks the target data reasonably well across the evaluated scenarios. While the temperature peak within the SSP5-3.4-over pathway is marginally underestimated, the broader trends are captured with sufficient accuracy.

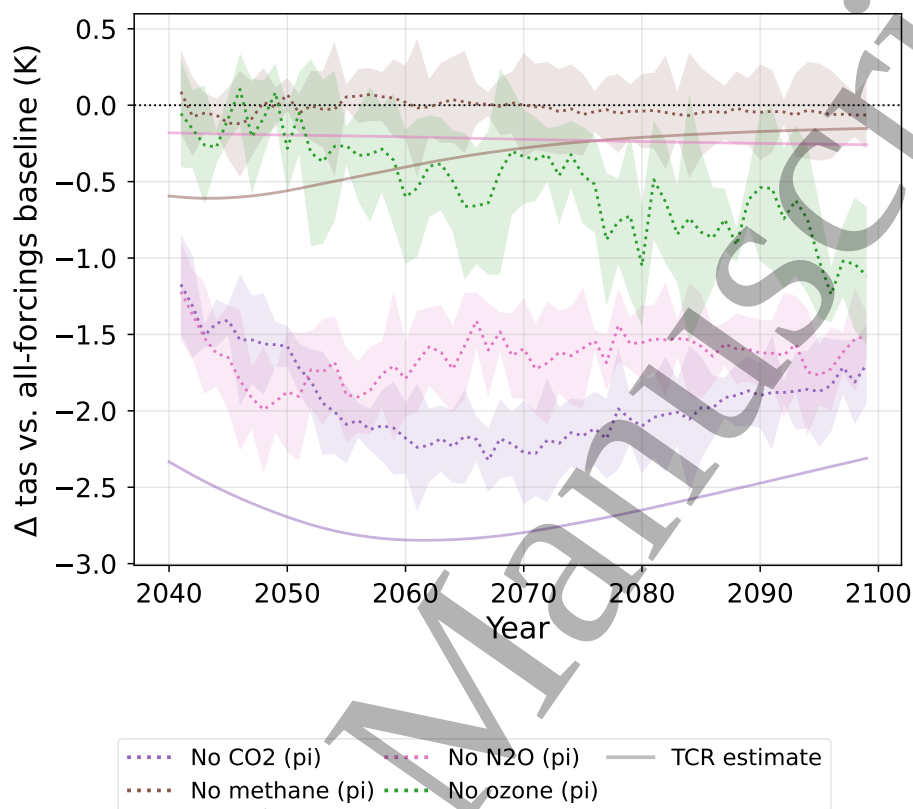


Figure 9: CanESM5 energy-score model, single-forcing piControl-held ablations under SSP5-3.4: yearly global-mean *tas* difference from the all-forcings baseline. Each forcing is set to pi-control levels and held for the entire rollout, the rest are left the same. CanESM5’s aerosol forcing is excluded from training entirely (see above), so no aerosol ablation exists here. For CO₂/CH₄/N₂O, the same-colored solid, semi-transparent line is that gas’s physically expected magnitude from the ERF formula scaled by CanESM5’s own TCR-derived sensitivity λ_{TCR} (TCR = 2.8K, section A.10); no comparable reference exists for ozone (section A.10). CanESM5’s single-forcing attribution behavior mirrors IPSL’s, suggesting that this behavior is a property of the training data and objective rather than of IPSL-CM6A-LR specifically, although only one further model is tested.

5 Discussion and Conclusion

We present AC-SSP, a deep-learning climate emulator capable of separating forced responses from internal climate variability while maintaining explicit spatial dynamics. In contrast to MESMER-M, AC-SSP resolves the full three-dimensional atmospheric profile. Furthermore, its generative capacity enables the production of diverse ensemble members from single initializations (e.g., IPSL trajectories), thus facilitating rigorous probabilistic analyses of deterministic scenarios.

We capture global trends for several SSP scenarios not seen at training time. We also show that the model has a reasonable forced response for the forcings included, although the impacts of CO₂ and, in particular, N₂O are overestimated. This suggests that decorrelating variables that covary in the historical record is important for robust generalization. This is reinforced by [80], who demonstrated that models trained on explicitly decorrelated data generalize far better to counterfactual scenarios. Expanding our training mixture with idealized point-source experiments—such as VolMIP [81] for stratospheric aerosol injections or ISA-MIP [82] for

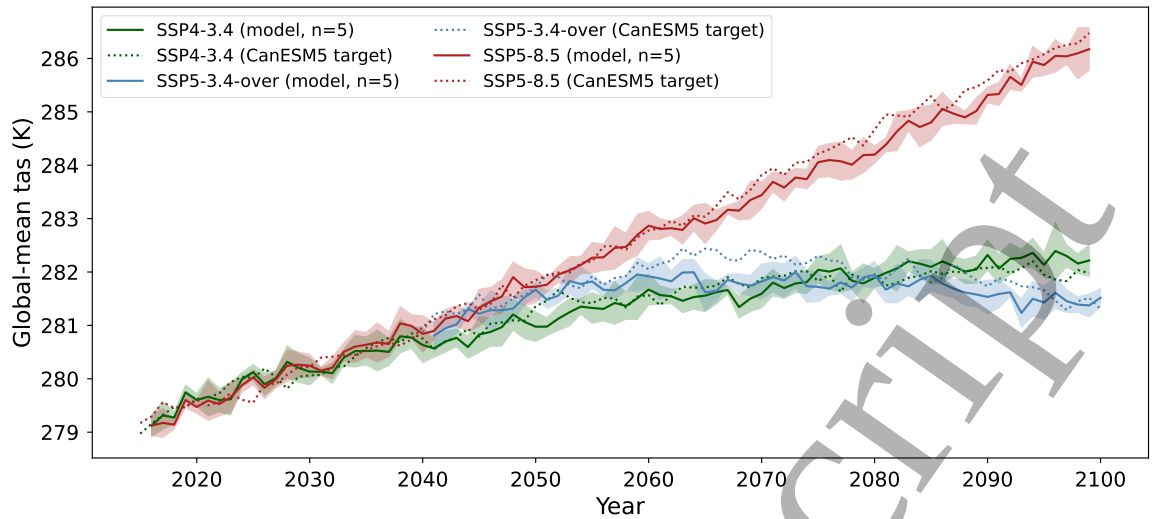


Figure 10: CanESM5 energy-score model, yearly global-mean *tas* (mean $\pm$ std band across ensemble members where applicable) for SSP4-3.4, SSP5-3.4, and SSP5-8.5 rollouts, each plotted against its own CanESM5 target realization. AC-SSP tracks all three CanESM5 scenarios reasonably well, including the SSP5-8.5 pathway, suggesting that the framework’s generalization ability is not specific to IPSL-CM6A-LR.

aerosol-climate interactions would further disentangle these signals. This could also be addressed through causal methods that could better learn the impact of the forcings on climate [83].

Because our training data intrinsically couples climate outputs with their generative forcings, we hypothesize that these prescribed forcings function primarily as attention-guiding mechanisms rather than wholly independent information sources. This is underlined by the fact that a model without forcings is numerically stable and lacks the information to reproduce a changing climate. This also supports the idea that our generative backbone can model internal variability without the support of forcings.

During these experiments, we saw a persistent bias in the Arctic region. Fixing this would plausibly require several changes beyond the scope of the current model: an area-weighted loss term to stop the tropics from dominating gradient updates; a native icosahedral or HEALPix grid representation, which does not artificially inflate polar spatial resolution and avoids the pole singularity altogether and using rotary positional embeddings (RoPE) in place of the backbone’s current positional encoding, which may generalize better to the extreme positions the polar regime effectively represents [84].

When evaluating model robustness, we distinguish two distinct forms of out-of-distribution behavior. As demonstrated in subsection 4.4, AC-SSP extrapolates competently to the global temperature trend of the SSP5-8.5 scenario, where the global CO₂ trajectory persistently exceeds the training domain. The other is the large initial increase of CO₂ seen with the abrupt-4xCO₂. This illuminates interesting questions about the role of the ocean in this model. The ocean, in nature, acts as a buffer and regulates and slows temperature [76]. We do not see any stopping forces in this model, as the abrupt-4xCO₂ generation quickly snaps to the expected equilibrium. It therefore may be necessary to include physical constraints, or include a more comprehensive representation of the ocean.

Ultimately, understanding irreducible scenario uncertainty is crucial for effective climate adaptation. We show that integrating forcings into a generative framework captures shifting atmospheric dynamics for held-out scenarios, including the policy-relevant overshoot trajectory SSP5-3.4, which makes AC-SSP a useful step toward exploring diverse human pathways. The abrupt-4xCO₂ experiment shows that the transient response remains a limitation. These findings provide a methodological basis for computationally efficient, probabilistic emulators for long-term climate projection, which could support the rapid scenario generation that policymakers need to assess how mitigation strategies alter centennial climate outcomes.

References

A Appendix

A.1 Climate Model Data

Table 5: Overview of IPSL scenario runs considered in the training and testing regime of AC-SSP. The scenario name and description identify the type of experiment run by IPSL; AC-SSP trains and tests on the outputs of a subset of those runs. All SSP runs are part of the ScenarioMIP experiments. “Ens. Mem.” stands for the number of ensemble members, i.e., repeated runs of the same scenario experiment, available. All runs provide monthly data for the years indicated. SSP4-3.4 is held out for validation, Abrupt-4xCO2 and SSP5-8.5 are held out for OOD validation, SSP5-3.4 is held out for testing, while all other scenarios and their ensembles are used for training.

Scenario Name	Description	Ens. Mem.	Years
<i>piControl</i>	Pre-industrial control simulation, used as a baseline with constant pre-industrial conditions	1	1850, repeated for 150 years
<i>historical</i>	Simulation of the climate using observed historical forcings	10	1850-2014
<i>SSP1-1.9</i>	A scenario where the world shifts toward sustainable development and net-zero emissions, successfully limiting global warming to approximately 1.5°C by 2100	4	2015-2100
<i>SSP1-2.6</i>	Future scenario following a low-emissions pathway (SSP1-2.6), aiming for strong mitigation.	5	2015-2100
<i>SSP2-4.5</i>	Intermediate emissions scenario (SSP2-4.5), representing moderate mitigation and socioeconomic development	6	2015-2100
<i>SSP3-7.0</i>	High emissions scenario (SSP3-7.0), reflecting regional rivalry and limited mitigation	10	2015-2100
<i>SSP4-3.4</i>	Low-emissions mitigation scenario (3.4 W m ⁻² by 2100) under SSP4, focusing on high inequality.	1	2015-2100
<i>SSP4-6.0</i>	A scenario characterized by high socioeconomic inequality and regional stratification, resulting in moderate climate policy success and a warming of roughly 2.7°C by 2100.	6	2015-2100
<i>SSP5-3.4</i>	SSP5-3.4 (Overshoot): A fossil-fuel-driven scenario that initially follows a high-emission path (SSP5-8.5) until 2040, followed by an aggressive, late-century shift to carbon removal to bring forcing down to 3.4 W/m ²	1	2040-2100
<i>SSP5-8.5</i>	Very high emissions scenario (SSP5-8.5), representing a “business-as-usual” pathway with minimal mitigation	4	2015-2100
<i>Abrupt-4xCO2</i>	An idealized scenario that begins with a 4x step of CO ₂ concentrations, starting from initial conditions at 1850	1	1850, repeated for 100 years

Table 6: Variables used in AC-SSP. Variables are categorized by Earth system component: Surface, Ocean, and Atmosphere. Each entry includes the CMIP-standardized name and a corresponding description. For oceanic and atmospheric components, the model incorporates all defined vertical depths and levels, respectively. All listed variables are predicted autoregressively.

Category	Variable	Long Name
Surface	<i>vas</i>	Northward Near-Surface Wind
	<i>uas</i>	Eastward Near-Surface Wind
	<i>pr</i>	Precipitation
	<i>ps</i>	Surface air pressure
	<i>tas</i>	Near-surface air temperature
	<i>evspsbl</i>	Evaporation including sublimation and transpiration
	<i>net_surface_flux</i>	Total positive downward flux between ocean and atmosphere
	<i>psl</i>	Sea level pressure
Ocean	<i>Depths (m): 0.51, 3.86, 8.09, 13.99, 22.76, 35.74, 53.85, 77.61, 108.03, 147.41</i>	
	<i>thetao</i>	Ocean temperature

Continued on next page

Table 6 – continued from previous page

Category	Variable	Long Name
Atmospheric	<i>Levels (hPa): 1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, 70, 50, 30, 20, 10</i>	
	<i>hus</i>	Specific humidity
	<i>ta</i>	Air temperature
	<i>ua</i>	Eastward wind
	<i>va</i>	Northward wind
	<i>zg</i>	Geopotential height

A.2 Forcing Dropout Logic Example

Here we provide a concrete example of the Forcing Dropout logic used in AC-SSP, introduced in section 3.3. On the remaining $1-p_{\text{all}}$ samples, we drop the forcings with a probability of p_{drop} . The two parameters are therefore only meaningful jointly: at $p_{\text{all}} = 1$, p_{drop} has no effect at all, since the masking branch is never reached; at $p_{\text{all}} = 0$, the model’s forcing exposure is controlled by p_{drop} . This allows us to ensure with some explicit probability how much the model sees the full forcing (p_{all}). For example, if we have $p_{\text{all}} = 0.8$ and $p_{\text{drop}} = 0.2$, then the model will see all of the forcings 80% of the time; on the remaining 20% of steps, each forcing independently has a 20% chance of being dropped: $P(\text{drop}) = (1 - p_{\text{all}}) \cdot p_{\text{drop}} = 0.2 \times 0.2 = 0.04$. With these parameters, any given forcing is dropped 4% of the time during training.

A.3 Stability Across Random Seeds (Energy-Score Model)

We check the stability of the energy-score model used throughout this paper: three otherwise-identical training runs of AC-SSP differing only in training-time random seed, all at step 22000, evaluated as 5-member ensembles on SSP4-3.4.

Scenario	Model	RMSE (K)	Spatial Std. Deviation	Ensemble Spread	Inter-annual Variability
SSP4-3.4	seed1	1.0356	13.4601	0.4600	0.6160
	seed2	1.1871	13.4339	0.5373	0.6283
	seed3	1.2591	13.6251	0.5872	0.6355
	MESMER-M	0.9390	13.2717	0.1179	0.7685
	IPSL	n/a	13.2520	n/a	0.6337

Table 7: Per-metric comparison of across three training seeds, step-22000, 5-member ensemble each, SSP4-3.4.

This energy-score model does not converge tightly across seeds. RMSE ranges from 1.0356K (seed1) to 1.2591K (seed3), a 22% relative spread between the best and worst seed. We do not have an explanation for why the three seeds order themselves so consistently across RMSE, ensemble spread and IAV rather than showing the seed-independent scatter we would expect from pure initialization noise. One possibility is that the energy-score loss’s added stochasticity (subsubsection 3.4.1) produces seed-unstable training dynamics. We note this as a limitation of the results reported for this model throughout this paper.

A.4 Forcing All-Present Probability at the Extremes: $p_{\text{all}} = 0.8$ vs. 1.0

subsubsection 4.1.2 finds $p_{\text{all}} = 0.8$ and $p_{\text{all}} = 1.0$ close to each other on aggregate SSP4-3.4 accuracy, both markedly better than lower p_{all} values. Here we compare their single-forcing piControl-held attribution behavior directly, SSP4-3.4, using the TCR-derived-sensitivity reference as Figure 4. Despite their similar aggregate accuracy, the two differ noticeably in attribution behavior (Figure 11, Figure 12): at $p_{\text{all}} = 1.0$, methane’s ablation response has the wrong sign relative to its TCR reference (positive rather than the expected negative), and N₂O’s over-attribution is more severe than at $p_{\text{all}} = 0.8$. Since $p_{\text{all}} = 1.0$ means the model never trains with any forcing masked, we read this as evidence that some exposure to forcing-dropout during training, not just high aggregate accuracy on the fully-forced case, is needed for reliable single-forcing attribution behavior at inference.

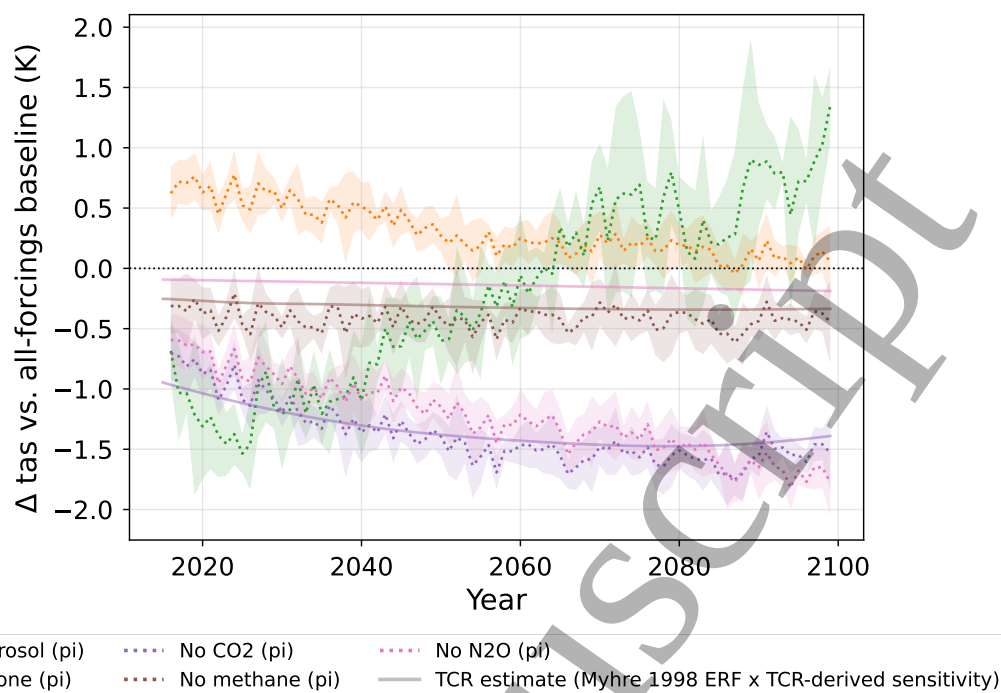


Figure 11: $p_{\text{all}} = 0.8$: single-forcing piControl-held ablations under SSP4-3.4, same methodology as Figure 4. At $p_{\text{all}} = 0.8$, methane's and aerosol's ablation traces have the correct sign and track their references closely, whereas ozone's ablation trace cools until about 2050 and then turns to warming (the wrong sign) by 2100, with a large ensemble spread.

A.5 Ensemble Size, Spread, and Skill (Energy-Score Model)

We report ensemble spread and RMSE for several ensemble sizes evaluated globally on SSP4-3.4 (val) with a much larger, 20-independently-seeded-rollout ensemble. We report both RMSE of the ensemble-mean prediction and the spread-skill-ratio (SSR), defined as

$\text{SSR} = \sqrt{\frac{M+1}{M}} \bar{\text{Var}}_M / \text{RMSE}$, where $\bar{\text{Var}}_M$ is the finite-sample-corrected ensemble variance averaged over all grid cells and the full rollout, and M is the ensemble size; a perfectly calibrated ensemble has $\text{SSR} = 1$, values below 1 indicate an under-dispersed (overconfident) ensemble, and values above 1 indicate an over-dispersed ensemble.

Ensemble size M	SSR	RMSE (K)	Spread (K)
1	n/a	2.525	n/a
5	1.012	2.544	2.576
10	1.048	2.432	2.550
15	1.071	2.405	2.576
20	1.077	2.388	2.573

Table 8: Global (area-weighted) *tas* SSR and RMSE of the ensemble-mean prediction as a function of ensemble size, for AC-SSP, SSP4-3.4 (val), full 85-year rollout, 20 independently-seeded rollouts of the same checkpoint. RMSE at $M = 1$ is not a meaningful spread comparison (a single member has no ensemble variance), so SSR/spread are reported as n/a there.

This energy-score model's ensemble is close to well-calibrated globally, and calibration steadily decreases as members are added rather than plateauing early: SSR rises from 1.012 at $M = 5$ to 1.077 at $M = 20$, starting close to the ideal value of 1.0 and drifting to a mild over-dispersion ($\text{SSR} \approx 1.05$ -1.08) beyond about 10 members. RMSE improves monotonically and substantially with ensemble size, from 2.544K at $M = 5$ down to 2.388K at $M = 20$, continuing to fall noticeably even past 10 members rather than saturating. This steady RMSE improvement indicates that a substantial share of this model's error is genuine sampling noise from the energy-score stochastic rollout rather than a shared bias, so larger ensembles are an effective way to reduce error for this model.

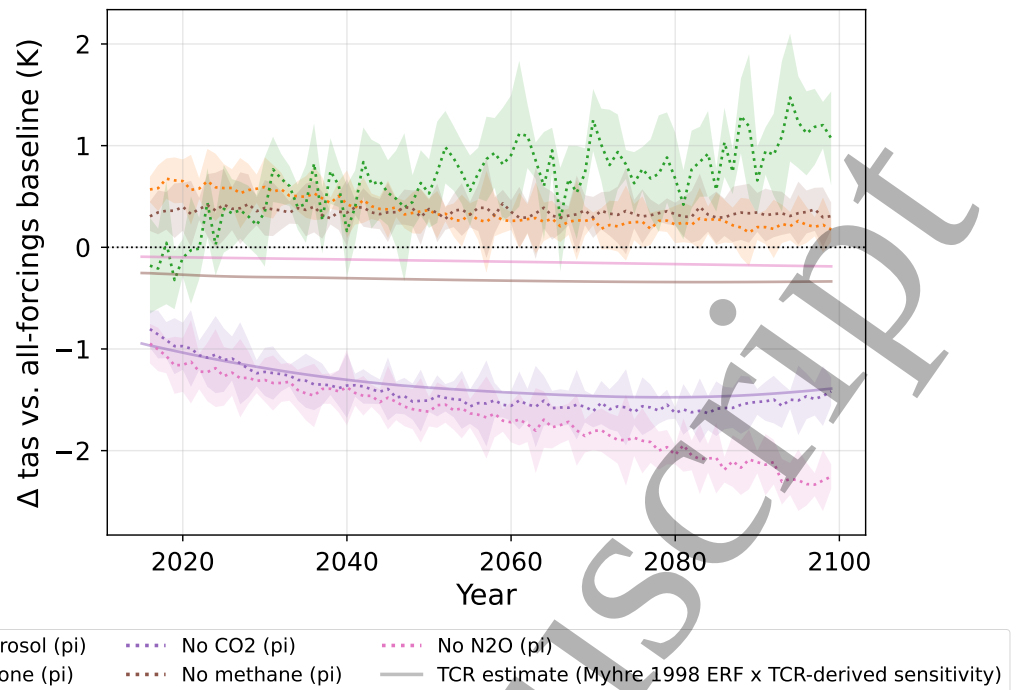


Figure 12: $p_{\text{all}} = 1.0$: single-forcing piControl-held ablations under SSP4-3.4, same methodology as Figure 4. At $p_{\text{all}} = 1.0$ (never trained with forcing-dropout), methane’s ablation trace has the *wrong sign* relative to its TCR reference, N₂O’s over-attribution is worse than at $p_{\text{all}} = 0.8$, and ozone’s ablation trace is warming (the wrong sign) for almost the whole rollout – always seeing the full forcing set during training leaves the model’s single-forcing-ablation behavior less reliable at inference, even though aggregate SSP4-3.4 accuracy is similar between the two (subsubsection 4.1.2).

A.6 Energy-Score Noise Scale: Effect on Calibration and Skill

The energy-score training objective (section 3.4.1) draws its stochasticity from Gaussian noise injected at inference time. Because this knob directly rescales the spread of the per-member stochastic perturbation without retraining, it offers a cheap way to trade ensemble spread against ensemble-mean RMSE. Table 9 reports SSR and RMSE of the 5-member ensemble-mean prediction for AC-SSP as the noise level is swept over $\{1.0, 0.8, 0.6, 0.5, 0.4\}$, on SSP4-3.4 (val), full 85-year rollout.

Noise scale	SSR	RMSE (K)	Spread (K)
1.0	1.012	2.544	2.576
0.8	0.913	1.965	1.794
0.6	0.711	1.967	1.399
0.5	0.601	1.972	1.185
0.4	0.485	1.981	0.961

Table 9: Global (area-weighted) *tas* SSR and RMSE of the 5-member ensemble-mean prediction as a function of `inference.energy_score_noise_scale`, for AC-SSP, full 85-year rollout.

Reducing the noise scale below 1.0 lowers SSR steadily. RMSE drops sharply between 1.0 and 0.8 (2.544 to 1.965 K) and changes little thereafter (1.965 to 1.981 K). In this configuration, then, noise scale is a trade-off between variance and accuracy: shrinking it trades away calibration for RMSE gains.

A.7 Flow Matching Performance of the Forced Response

In preliminary investigations, we adopted the residual flow-matching paradigm proposed by [1, 13]. This involved training a deterministic base model (utilizing the identical forcing dropout and pushforward mechanisms described in our current energy-score configuration) and subsequently training a flow-matching model to predict its step-to-step residuals. However, this approach yielded a marked degradation in the fidelity of the forced response, as illustrated in Figure 13.

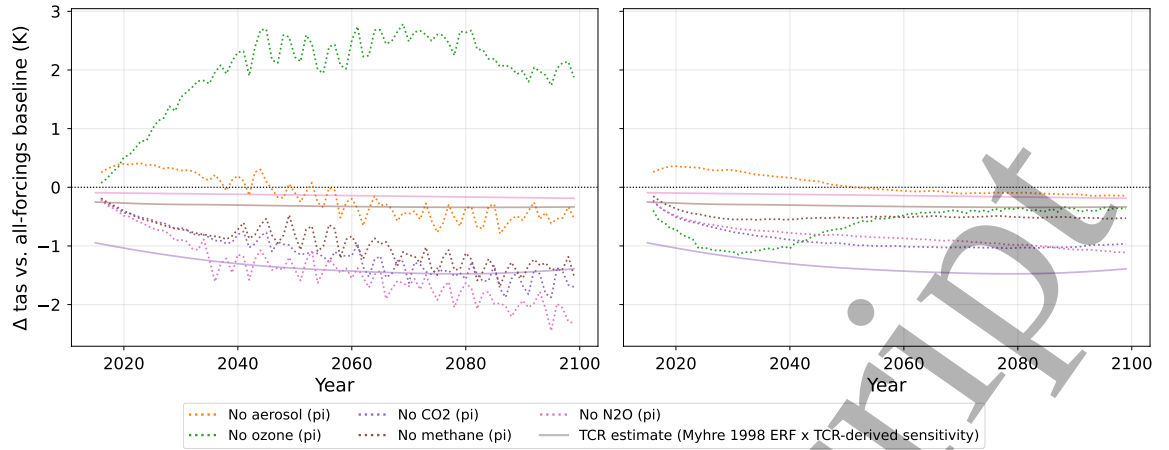


Figure 13: Single-forcing piControl-held ablations for SSP4-3.4: annual global-mean *tas* deviation from each model’s own all-forcings baseline, for all five single-forcing ablations. The flow-matching model (left) is compared against its underlying deterministic base model (right). For CO₂/methane/N₂O, the same-colored solid, semi-transparent line is that gas’s TCR-derived sensitivity reference (same convention as Figure 4); no aerosol/ozone reference is shown, as the simple ERF formulas used here do not cover those forcings. The deterministic base model’s ablation traces evolve smoothly and track their TCR references, while the flow-matching model’s traces are dominated by high-frequency, step-to-step noise that obscures the underlying forced-response signal, motivating the switch to the energy-score objective.

Applying our piControl-held ablation methodology to both architectures reveals a stark contrast. The deterministic base model (right column) produces traces that align closely with their transient climate response (TCR) references, evolving smoothly throughout the rollout. Conversely, the flow-matching model (left column) introduces pronounced high-frequency, step-to-step noise. While it tracks the broad underlying trend without runaway divergence—indicating that the forced-response signal is partially preserved rather than entirely obliterated—the variance is severely inflated. Notably, the no-aerosol and no-ozone ablations exhibit the largest absolute deviations in both architectures; the flow model’s no-ozone trace, in particular, diverges by approximately 2.0 to 2.8 K by mid-century before stabilizing. We attribute this degradation to a misalignment of optimization objectives. The single-step flow-matching objective aggressively fits high-frequency, stochastic local variability, effectively overriding the low-amplitude, long-term forcing signal that the pushforward mechanism successfully embedded in the deterministic base model. Because extending flow matching to multi-step (pushforward) fine-tuning is computationally prohibitive, correcting this imbalance within the flow paradigm was unfeasible. We therefore required a generative framework natively compatible with multi-step optimization. Although recent literature has demonstrated multi-step fine-tuning via consistency models [85], our attempts to distill this specific flow model into a consistency model proved unsuccessful, ultimately necessitating the architectural pivot to the energy-score approach detailed in the main text.

A.8 Metrics

This section defines the diagnostic metrics used to evaluate model performance, variability, and uncertainty. Throughout these definitions, Y denotes the target (ground truth) values and $\hat{Y}$ denotes the predicted values.

Latitude-Weighted RMSE To account for the decreasing surface area of grid cells toward the poles, we define the Latitude-Weighted Root Mean Square Error (RMSE) for an ensemble member e as:

$$\text{RMSE}_e = \sqrt{\frac{\sum_{t,\phi,\lambda} \cos(\phi) \cdot (Y_{e,t,\phi,\lambda} - \hat{Y}_{e,t,\phi,\lambda})^2}{\sum_{t,\phi,\lambda} \cos(\phi)}} \quad (5)$$

where ϕ is the latitude, λ is the longitude, and t is the timestep. This value is then averaged across ensemble members.

NRMSE To allow for meaningful comparisons across variables with varying scales, we utilize the Normalized RMSE (NRMSE). This is calculated by applying the Latitude-Weighted RMSE to variables that have been pre-normalized by their respective standard deviations.

Interannual Variability (IAV) The grid-point IAV measures the magnitude of year-to-year fluctuations after removing the long-term forced trend. For an ensemble member e , it is defined as:

$$IAV_e(\phi, \lambda) = \sqrt{\frac{1}{N_t - 1} \sum_{t=1}^{N_t} \left(\hat{Y}_{e,t,\phi,\lambda} - \hat{\int}_{e,\phi,\lambda}(t) \right)^2} \quad (6)$$

where $\hat{\int}_{e,\phi,\lambda}(t)$ represents the least-squares linear fit of the time series at that specific grid point and time. Each IAV calculation is independently detrended to prevent long-term warming trends from conflating the assessment of annual variability. IAV is based on temperature anomalies relative to a 1994–2014 climatological baseline.

Ensemble Spread The ensemble spread quantifies the uncertainty between different model realizations. First, we calculate the global spatial average for each ensemble member e at timestep t :

$$\bar{Y}_{e,t} = \frac{1}{N_\phi N_\lambda} \sum_{\phi=1}^{N_\phi} \sum_{\lambda=1}^{N_\lambda} \hat{Y}_{e,t,\phi,\lambda} \quad (7)$$

Next, we compute the standard deviation across ensemble members (N_E) to find the instantaneous spread σ_t :

$$\sigma_t = \sqrt{\frac{1}{N_E - 1} \sum_{e=1}^{N_E} (\bar{Y}_{e,t} - \langle \bar{Y}_{\cdot,t} \rangle)^2} \quad (8)$$

where $\langle \bar{Y}_{\cdot,t} \rangle$ is the ensemble mean of the spatial averages. The final Average Ensemble Spread is the temporal mean of these values: $\frac{1}{N_t} \sum_t \sigma_t$.

Spatial Standard Deviation To evaluate the model's ability to capture spatial gradients, we calculate the standard deviation over the spatial dimensions (ϕ, λ) for each timestep, subsequently averaged over time:

$$\sigma_{spatial} = \frac{1}{N_t} \sum_{t=1}^{N_t} \left(\sqrt{\frac{1}{N_\phi N_\lambda - 1} \sum_{\phi,\lambda} (\hat{Y}_{t,\phi,\lambda} - \bar{Y}_t)^2} \right) \quad (9)$$

Hydrostatic Balance We compare two different ways of calculating the thickness of atmospheric layers in order to understand if the model is learning a good relationship between ta and zg . First, we check the Geometric Thickness, which is the direct difference between the geopotential height (zg) values at two adjacent pressure levels. Next, we check the Hydrostatic Thickness. This is the thickness calculated using the temperature (ta) and the known pressure levels. If the generated states are physically consistent, these two values should be nearly identical.

The fundamental relationship being checked is derived from the hydrostatic equation and the ideal gas law:

$$\Delta z = \frac{R_d \bar{T}}{g} \ln \left(\frac{p_{lower}}{p_{upper}} \right) \quad (10)$$

Where: R_d is the gas constant for dry air ($287.05 \text{ J kg}^{-1} \text{ K}^{-1}$). g is the gravitational acceleration (9.80665 m s^{-2}). $\bar{T}$ is the mean temperature of the layer. p is the pressure.

Layer Thickness from Geopotential (Δz_{geo}): This is calculated as the difference between adjacent levels in the geopotential height array zg :

$$\Delta z_{geo} = z_{g,i+1} - z_{g,i} \quad (11)$$

Average Temperature ($\bar{T}$): The mean layer temperature using the arithmetic mean of the temperature T at the boundaries:

$$\bar{T} = \frac{T_i + T_{i+1}}{2} \quad (12)$$

Log-Pressure Difference ($\Delta \ln p$): The difference in natural logarithms of pressure:

$$\Delta \ln p = \ln \left(\frac{p_i}{p_{i+1}} \right) \quad (13)$$

Hydrostatic Thickness (Δz_{hydro}): Combining these, the theoretical thickness based on the hydrostatic assumption is:

$$\Delta z_{hydro} = \left(\frac{R_d \bar{T}}{g} \right) \Delta \ln p \quad (14)$$

Finally, we calculate the residual (the error) and the Mean Absolute Error (MAE):

$$\varepsilon = \Delta z_{geo} - \Delta z_{hydro} \quad (15)$$

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{N} \sum_{n=1}^N |\varepsilon_n| \quad (16)$$

A.9 Energy Balance

Energy conservation is one of the few constraints that must hold for any closed representation of the climate system regardless of resolution or parameterization choice, and it is the standard lens through which the observed climate system's radiative and heat-storage terms are reconciled [73, 74]. The emulator represents only the atmosphere and the upper ocean; deep-ocean, land, and cryosphere heat storage are not included, so the budget is not closed. We adapt the same bookkeeping to check the energy-budget consistency of the emulator, i.e. whether the rate of change of the emulator's modeled energy reservoirs (atmosphere and upper ocean) is consistent with its own net surface flux.

A.9.1 Column atmospheric energy The column-integrated atmospheric energy per unit area combines internal (sensible), potential, kinetic, and latent contributions,

$$E_{\text{atm}} = \int_0^{ps} \left(c_p t a + g z g + \frac{1}{2} (u a^2 + v a^2) + L_v h u s \right) \frac{dp}{g}, \quad (17)$$

with $c_p = 1004.6 \text{ J kg}^{-1} \text{ K}^{-1}$ and $L_v = 2.5 \times 10^6 \text{ J kg}^{-1}$. This is from equation 3.14 in (<https://sci-hub.fr/10.1103/RevModPhys.56.365>). We discretize Eq. 17 on the fixed pressure grid by assigning each level i a layer $[u_i, \ell_i]$ bounded by the midpoints to its neighbors ($u_i = (p_{i-1} + p_i)/2$, $\ell_i = (p_i + p_{i+1})/2$, with $u_1 = 0$ at the model top and $\ell_n \rightarrow \infty$ at the surface-most level), and clip each layer to the actual surface pressure of the grid cell:

$$\Delta p_i = \max(0, \min(\ell_i, ps) - u_i), \quad E_{\text{atm}} \approx \sum_i \frac{e_i \Delta p_i}{g}, \quad (18)$$

where e_i is the bracketed integrand of Eq. 17 evaluated at level i . This clipping is necessary because ps varies with orography and surface pressure while the pressure levels are fixed: levels that lie below the actual surface ($u_i \geq ps$) are excluded, and the lowest valid layer is truncated at ps rather than at its nominal lower bound, following the standard treatment of terrain-following mass integrals in fixed-pressure-level archives.

A.9.2 Ocean heat content The upper-ocean heat content per unit area is the depth integral

$$E_{\text{oce}} = \int_0^{-H} \rho_w c_{pw} \text{thetao} dz, \quad (19)$$

with $\rho_w = 1026 \text{ kg m}^{-3}$ and $c_{pw} = 3991 \text{ J kg}^{-1} \text{ K}^{-1}$ [86], seawater density and seawater heat capacity [87] discretized analogously to Eq. 18 using layer thicknesses Δz_i derived from the midpoints between successive depth levels $z_1, \dots, z_m$ ($H = 147.4 \text{ m}$ here).

A.9.3 Global integration and the energy-budget consistency check Column energies and net flux are integrated over the sphere using the area element

$$A(\phi) = R_{\oplus}^2 \cos \phi \Delta \phi \Delta \lambda, \quad (20)$$

with $R_{\oplus} = 6.371 \times 10^6 \text{ m}$, on the model's regular 144×144 latitude-longitude grid. Writing $\mathbf{E}_{\text{atm}}(t) = \sum_{i,j} E_{\text{atm}}(t, i, j) A_{i,j}$ (and analogously for $\mathbf{E}_{\text{oce}}$ and $\mathbf{F}_{\text{net}}$), energy-budget consistency at monthly resolution requires, to first approximation,

$$\frac{\Delta \mathbf{E}_{\text{atm}} + \Delta \mathbf{E}_{\text{oce}}}{\Delta t} \approx \overline{\mathbf{F}_{\text{net}}}, \quad (21)$$

where Δt is taken as the average calendar month (365.25/12 days) and $\overline{\mathbf{F}_{\text{net}}}$ is the mean of $\mathbf{F}_{\text{net}}$ over the two months bounding Δt . The residual of Eq. 21 is reported both as a global time series and, for the trailing decade of a rollout, as a spatial map of $\langle E_{\text{atm}} \rangle$ and $\langle E_{\text{oce}} \rangle$ for the emulator, the reference, and their difference, which localizes systematic bias rather than only detecting it in the global mean.

A.10 Effective Radiative Forcing and Transient Climate Response Analysis

For the greenhouse gases (CO_2 , CH_4 , N_2O), we build a physically expected Δt_{as} reference for each ablation by combining two standard quantities, Effective Radiative Forcing (ERF) and Transient Climate Response (TCR), in three steps.

We convert each gas's concentration trajectory $C(t)$ into an effective radiative forcing $\text{ERF}(t)$ in W/m^2 , relative to its pre-industrial reference concentration C_0 (CO_2 : 278 ppm, CH_4 : 700 ppb, N_2O : 270 ppb). For CO_2 we use the standard closed-form expression of [88]:

$$\text{ERF}_{\text{CO}_2}(C) = 5.35 \ln(C/C_0). \quad (22)$$

For CH_4 and N_2O we instead use the expressions of [89] which additionally capture the cross-sensitivity of each gas's ERF to the other's concentration and to CO_2 :

$$\text{ERF}_{\text{CH}_4}(M) = (a_3 \bar{M} + b_3 \bar{N} + 0.043) (\sqrt{M} - \sqrt{M_0}) \quad (23)$$

$$\text{ERF}_{\text{N}_2\text{O}}(N) = (a_2 \bar{C} + b_2 \bar{N} + c_2 \bar{M} + 0.117) (\sqrt{N} - \sqrt{N_0}) \quad (24)$$

where $a_3 = -1.3 \times 10^{-6}$, $b_3 = -8.2 \times 10^{-6}$, $a_2 = -8.0 \times 10^{-6}$, $b_2 = 4.2 \times 10^{-6}$, and $c_2 = -4.9 \times 10^{-6}$ (all in W/m^2 per the relevant concentration unit, ppm for CO_2 and ppb for $\text{CH}_4/\text{N}_2\text{O}$); the other two gases are held at their own current-trajectory values (not pre-industrial) in each formula, following [89].

ERF alone is a unit of energy flux and turning it into an expected Δt_{as} requires a climate sensitivity. We use TCR (Transient Climate Response: the transient warming expected at CO_2 doubling under a 1%/yr ramp, [90]). Rather than a single generic TCR value, we use a model-specific estimate for each source ESM: $\text{TCR} = 2.45\text{K}$ for IPSL-CM6A-LR and $\text{TCR} = 2.8\text{K}$ for CanESM5 [14, 79]. Because TCR is defined per unit of CO_2 -doubling forcing, we convert it to a sensitivity in K per W/m^2 by dividing by the forcing at CO_2 doubling, $F_{2 \times \text{CO}_2} = 5.35 \ln 2 \approx 3.71 \text{W}/\text{m}^2$ (the same constant appearing in the CO_2 ERF formula above, evaluated at $C/C_0 = 2$):

$$\lambda_{\text{TCR}} = \text{TCR}/F_{2 \times \text{CO}_2}. \quad (25)$$

Multiplying a gas's ERF trajectory by λ_{TCR} converts it into the physically expected transient temperature response to that gas alone, under the linear-response assumption TCR itself already embeds:

$$\Delta t_{\text{as,expected}}(t) = -\lambda_{\text{TCR}} \cdot \text{ERF}_{\text{gas}}(C_{\text{gas}}(t)), \quad (26)$$

where the sign is negative because setting a gas to its pre-industrial control level removes its forcing relative to the all-forcings baseline, so a gas with positive (warming) ERF is expected to produce cooling once held at pi-control. This is the solid reference trace plotted against each model's own piControl-held ablation trajectory (dotted) seen throughout this paper.

Unlike a per-gas TCR fit to a specific single-forcing experiment (only available from the INCA-driven IPSL DAMIP suite), the $\text{CH}_4/\text{N}_2\text{O}$ $\text{ERF} \times \text{TCR}$ reference above needs only each gas's own concentration trajectory (available for any model from the same input4MIPs forcing files used at training time) and the host model's own overall TCR; we therefore construct it identically for both IPSL-CM6A-LR and CanESM5, using [89]'s $\text{CH}_4/\text{N}_2\text{O}$ ERF formulas and each model's own TCR (2.45K for IPSL-CM6A-LR, 2.8K for CanESM5 [14, 79]).

A.11 CanESM5 vs. IPSL Training Data

Here we explain the data and processing for the CanESM5 data, as well as a comparison to available data of the IPSL. We retain CanESM5 in its native 64×128 grid, in contrast to the 144×144 configuration of the IPSL model. Consequently, we apply an altered window size in the SWIN component ([8, 16] rather than [6, 10]). CanESM5 provides 19 pressure levels, which constitutes a superset of the 17 levels present in IPSL. Both models share the [100000, ..., 1000] Pa range, but CanESM5 includes two additional uppermost levels (500 Pa and 100 Pa) that we cut. For the ocean temperature field (*thetao*), CanESM5 uses 45 levels beginning at the surface, whereas IPSL employs 10 levels starting at a depth of 0.51 m. To maintain layer-integrated heat content across these differing discretizations, we project the CanESM5 data onto the IPSL target depths using a conservative, bounds-overlap-weighted vertical remapping. The derivation of forcing fields remains identical across both models, except for the aerosol forcings. Because the aerosol forcing in CanESM5 is structured differently from the INCA-derived fields used for IPSL (detailed in subsection 3.1), we omit the aerosol forcing for CanESM5.

The datasets also exhibit discrepancies in ensemble size, as the number of realizations for CanESM5 is smaller than that of IPSL. Table 10 (subsection A.11) shows the realizations utilized for each shared training experiment. Additionally, two specific scenarios present in the IPSL training, SSP1-1.9 and SSP4-6.0, lack CanESM5 equivalents within our dataset. Table 10 gives the per-experiment realization counts referenced in subsection 4.5.

Experiment	IPSL realizations	CanESM5 realizations
historical	10	10
piControl	2	0 (excluded)
abrupt-4xCO2	1	0 (excluded)
SSP1-1.9	6	0 (excluded)
SSP1-2.6	6	3
SSP2-4.5	6	5
SSP3-7.0	10	9
SSP4-3.4 (validation target)	1	1
SSP4-6.0	6	0 (excluded)
SSP5-3.4-over (test target)	1	1
SSP5-8.5 (OOD validation target)	6	2

Table 10: Number of ensemble-member realizations (CMIP6 *r*ip1f1* variants) used per training/evaluation experiment, IPSL vs. CanESM5. Counts reflect what our processed archive actually contains, not necessarily the full set nominally available from each model’s CMIP6 submission. CanESM5’s historical realizations are a non-contiguous subset (r1, r2, r3, r5, r7, r10, r11, r12, r13, r14, r16, r18, r19, r20, r22, r24, r25), unlike IPSL’s contiguous r1-r10.

A.12 Rest of Model Performance

For completeness, Table 11 and Table 14 report per-level performance for every remaining atmospheric and ocean variable not otherwise broken out elsewhere in this paper: last-decade global RMSE against the CMIP target, spatial standard deviation (SpStd), and detrended inter-annual variability (IAV), for AC-SSP on SSP5-3.4 with a 5-member ensemble. SpStd and IAV cells are given as *model (IPSL)*, so the model’s own value can be compared directly against the IPSL target’s at the same level. Spatial std and IAV are computed identically to the whole-column (*tas*) diagnostics used elsewhere in this paper (subsection A.5), applied per pressure/depth level over the same final-decade window as the RMSE figures.

Acknowledgments

The authors are grateful to the CLEPS infrastructure from the Inria of Paris for providing resources and support. To process the data from IPSL, this study benefited from the IPSL mesocenter ESPRI facility, which is supported by CNRS, UPMC, Labex L-IPSL, CNES and Ecole Polytechnique.

Funding

G. Clyne and C. Monteleoni were primarily supported by the Choose France Chair in AI, from the French government.

Data availability

The code for this project is available at <https://github.com/grahamclyne/ArchesClimate-SSP>. The model and a data sample is available at <https://huggingface.co/gclyne/ArchesClimate> and <https://huggingface.co/datasets/gclyne/ArchesClimate-data>. With the exception of the aerosol forcings, the training data can be found at <https://esgf-metagrid.cloud.dkrz.de/search>.

References

[1] Clyne G, Couairon G, Gastineau G, Monteleoni C, Charantonis A. ArchesClimate: Probabilistic Decadal Ensemble Generation With Flow Matching. arXiv; 2025.

[2] Pörtner HO, Roberts DC, Adams H, Adelekan I, Adler C, Adrian R, et al. Technical Summary. In: Pörtner HO, Roberts DC, Tignor MMB, Poloczanska ES, Mintenbeck K, Alegría A, et al, editors. Climate Change 2022: Impacts, Adaptation and Vulnerability.

Pressure level (hPa)	RMSE	ta [K]		RMSE	zg [m]	
		SpStd	IAV		SpStd	IAV
1000	0.373	11.6 (12.4)	0.441 (0.463)	4.04	53.1 (56.2)	5.28 (6.78)
925	0.307	11.4 (11.5)	0.513 (0.442)	3.17	95.7 (96.1)	6.98 (8.54)
850	0.327	10.9 (10.9)	0.499 (0.436)	3.09	117 (117)	7.52 (8.62)
700	0.250	12.5 (12.6)	0.443 (0.429)	3.61	177 (176)	8.35 (10.7)
600	0.250	11.5 (11.5)	0.444 (0.403)	4.19	233 (232)	8.62 (12.0)
500	0.216	11.8 (11.7)	0.458 (0.399)	4.55	293 (292)	10.8 (13.7)
400	0.219	11.8 (11.8)	0.378 (0.394)	6.77	368 (366)	11.8 (15.7)
300	0.221	10.8 (10.7)	0.297 (0.369)	8.51	463 (461)	14.5 (18.2)
250	0.183	9.03 (8.98)	0.340 (0.399)	8.45	516 (513)	16.0 (19.4)
200	0.233	5.31 (5.33)	0.532 (0.543)	8.86	561 (558)	15.6 (20.4)
150	0.271	4.09 (4.19)	0.564 (0.636)	9.55	573 (571)	15.4 (21.7)
100	0.271	9.33 (9.30)	0.496 (0.582)	11.5	528 (526)	15.1 (24.5)
70	0.302	8.38 (8.36)	0.502 (0.661)	11.8	478 (479)	15.9 (26.6)
50	0.320	5.57 (5.62)	0.491 (0.678)	11.3	462 (465)	18.6 (28.7)
30	0.312	3.52 (3.74)	0.453 (0.617)	13.2	480 (487)	22.2 (33.1)
20	0.300	3.07 (3.35)	0.456 (0.627)	16.2	511 (520)	25.4 (36.4)
10	0.300	2.41 (2.62)	0.483 (0.586)	18.3	559 (574)	28.1 (41.2)

Table 11: *AC-SSP* atmospheric-column performance by pressure level, SSP5-3.4-over, last decade of the rollout, for *ta* and *zg*. SpStd/IAV given as model (IPSL). Continued for *hus/ua* in Table 12 and *va* in Table 13.

Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press; 2022. p. 37-118.

[3] Climate-Resilient Pathways: Adaptation, Mitigation, and Sustainable Development. In: Field CB, Barros VR, Dokken DJ, Mach KJ, Mastrandrea MD, editors. Climate Change 2014 Impacts, Adaptation, and Vulnerability. Cambridge: Cambridge University Press; 2014. p. 1101-31.

[4] Chen D, Rojas M, Samset BH, Cobb K, Diongue-Niang A, Edwards P, et al. Framing, Context, and Methods. In: Masson-Delmotte V, Zhai P, Pirani A, Connors SL, Péan C, Berger S, et al, editors. Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press; 2021. p. 147-286.

[5] Manabe S, Wetherald RT. Thermal Equilibrium of the Atmosphere with a Given Distribution of Relative Humidity. Journal of the Atmospheric Sciences. 1967 May;24(3):241-59.

[6] Manabe S, Bryan K. Climate Calculations with a Combined Ocean-Atmosphere Model. Journal of the Atmospheric Sciences. 1969 Jul;26(4):786-9.

[7] Eyring V, Bony S, Meehl GA, Senior CA, Stevens B, Stouffer RJ, et al. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) Experimental Design and Organization. Geoscientific Model Development. 2016;9(5):1937-58.

[8] O'Neill BC, Tebaldi C, van Vuuren DP, Eyring V, Friedlingstein P, Hurtt G, et al. The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6. Geoscientific Model Development. 2016 Sep;9(9):3461-82.

[9] van Vuuren DP, Riahi K, Calvin K, Dellink R, Emmerling J, Fujimori S, et al. The Shared Socio-economic Pathways: Trajectories for Human Development and Global Environmental Change. Global Environmental Change. 2017 Jan;42:148-52.

[10] Meinshausen M, Nicholls ZRJ, Lewis J, Gidden MJ, Vogel E, Freund M, et al. The Shared Socio-Economic Pathway (SSP) Greenhouse Gas Concentrations and Their Extensions to 2500. Geoscientific Model Development. 2020 Aug;13(8):3571-605.

Pressure level (hPa)	RMSE	hus [kg/kg]		RMSE	ua [m/s]	
		SpStd	IAV		SpStd	IAV
1000	1.30×10^{-4}	0.00538 (0.00544)	2.05×10^{-4} (2.20×10^{-4})	0.199	3.97 (3.94)	0.445 (0.535)
925	1.21×10^{-4}	0.00477 (0.00475)	1.87×10^{-4} (1.90×10^{-4})	0.246	5.11 (5.12)	0.640 (0.666)
850	1.03×10^{-4}	0.00386 (0.00385)	1.75×10^{-4} (1.73×10^{-4})	0.269	5.24 (5.29)	0.652 (0.713)
700	8.26×10^{-5}	0.00222 (0.00221)	1.28×10^{-4} (1.40×10^{-4})	0.337	5.84 (5.87)	0.797 (0.843)
600	5.92×10^{-5}	0.00155 (0.00153)	9.18×10^{-5} (1.09×10^{-4})	0.349	6.39 (6.44)	0.883 (0.937)
500	3.71×10^{-5}	7.97×10^{-4} (7.92×10^{-4})	5.54×10^{-5} (6.48×10^{-5})	0.414	7.23 (7.27)	1.13 (1.09)
400	1.85×10^{-5}	3.84×10^{-4} (3.80×10^{-4})	2.42×10^{-5} (2.94×10^{-5})	0.483	8.47 (8.49)	1.18 (1.30)
300	6.74×10^{-6}	1.38×10^{-4} (1.36×10^{-4})	7.39×10^{-6} (9.30×10^{-6})	0.542	10.4 (10.4)	1.35 (1.51)
250	3.18×10^{-6}	6.54×10^{-5} (6.42×10^{-5})	3.21×10^{-6} (4.09×10^{-6})	0.567	11.7 (11.7)	1.36 (1.59)
200	1.14×10^{-6}	2.33×10^{-5} (2.28×10^{-5})	1.09×10^{-6} (1.42×10^{-6})	0.559	12.9 (12.9)	1.24 (1.58)
150	2.73×10^{-7}	5.34×10^{-6} (5.21×10^{-6})	2.59×10^{-7} (3.65×10^{-7})	0.505	12.8 (12.8)	1.07 (1.49)
100	2.38×10^{-8}	4.04×10^{-7} (3.93×10^{-7})	2.94×10^{-8} (4.09×10^{-8})	0.465	10.3 (10.4)	1.03 (1.40)
70	9.54×10^{-9}	7.89×10^{-8} (7.32×10^{-8})	9.39×10^{-9} (1.30×10^{-8})	0.532	10.3 (10.3)	0.968 (1.32)
50	8.27×10^{-9}	7.43×10^{-8} (7.17×10^{-8})	7.76×10^{-9} (1.29×10^{-8})	0.660	11.0 (11.1)	1.11 (1.69)
30	9.81×10^{-9}	4.03×10^{-8} (3.58×10^{-8})	6.98×10^{-9} (1.08×10^{-8})	0.813	12.3 (12.8)	1.43 (2.11)
20	9.41×10^{-9}	1.42×10^{-8} (7.47×10^{-9})	6.52×10^{-9} (8.48×10^{-9})	0.863	13.7 (14.2)	1.76 (2.41)
10	8.31×10^{-9}	8.72×10^{-9} (4.00×10^{-9})	6.33×10^{-9} (8.66×10^{-9})	1.96	14.0 (15.1)	2.19 (3.30)

Table 12: *AC-SSP* atmospheric-column performance by pressure level, SSP5-3.4-over, last decade of the rollout, for *hus* and *ua*. SpStd/IAV given as model (IPSL). Continued from Table 11.

Pressure level (hPa)	RMSE	va [m/s]	
		SpStd	IAV
1000	0.158	1.81 (1.82)	0.375 (0.409)
925	0.187	1.90 (1.89)	0.466 (0.459)
850	0.196	1.62 (1.63)	0.460 (0.447)
700	0.212	1.77 (1.80)	0.518 (0.520)
600	0.233	1.72 (1.75)	0.580 (0.586)
500	0.275	1.67 (1.76)	0.558 (0.689)
400	0.320	1.90 (1.93)	0.781 (0.829)
300	0.373	2.22 (2.21)	1.05 (0.977)
250	0.391	2.31 (2.38)	0.957 (1.03)
200	0.397	2.34 (2.48)	0.751 (1.00)
150	0.365	2.22 (2.37)	0.600 (0.873)
100	0.290	1.90 (2.08)	0.521 (0.709)
70	0.257	1.87 (2.10)	0.451 (0.613)
50	0.264	2.12 (2.43)	0.451 (0.618)
30	0.319	2.69 (3.08)	0.504 (0.709)
20	0.367	3.14 (3.59)	0.545 (0.817)
10	0.450	3.68 (4.18)	0.624 (0.987)

Table 13: *AC-SSP* atmospheric-column performance by pressure level, SSP5-3.4-over, last decade of the rollout, for *va*. SpStd/IAV given as model (IPSL). Continued from Table 12.

[11] Maher N, Power SB, Marotzke J. More Accurate Quantification of Model-to-Model Agreement in Externally Forced Climatic Responses over the Coming Century. *Nature Communications*. 2021 Feb;12(1):788.

[12] Lehner F, Deser C, Maher N, Marotzke J, Fischer EM, Brunner L, et al. Partitioning Climate Projection Uncertainty with Multiple Large Ensembles and CMIP5/6. *Earth System Dynamics*. 2020 May;11(2):491-508.

[13] Couairon G, Singh R, Charantonis A, Lessig C, Monteleoni C. ArchesWeatherGen: Skillful and Compute-Efficient Probabilistic Weather Forecasting with Machine Learning. *Science Advances*. 2026 Apr;12(17):eadx2372.

[14] Boucher O, Servonnat J, Albright AL, Aumont O, Balkanski Y, Bastrikov V, et al. Presentation and Evaluation of the IPSL-CM6A-LR Climate Model. *Journal of Advances in Modeling Earth Systems*. 2020;12(7):e2019MS002010.

[15] Nath S, Lejeune Q, Beusch L, Seneviratne SI, Schleussner CF. MESMER-M: An Earth System Model Emulator for Spatially Resolved Monthly Temperature. *Earth System Dynamics*. 2022 Apr;13(2):851-77.

Depth level (m)	RMSE	thetao [K]	
		SpStd	IAV
0.5	0.255	11.1 (11.2)	0.259 (0.363)
3.9	0.253	11.1 (11.2)	0.250 (0.360)
8.1	0.254	11.1 (11.2)	0.243 (0.359)
14.0	0.264	11.2 (11.3)	0.241 (0.358)
22.8	0.283	11.3 (11.3)	0.245 (0.361)
35.7	0.309	11.0 (11.0)	0.240 (0.363)
53.9	0.328	10.5 (10.5)	0.238 (0.368)
77.6	0.348	9.80 (9.81)	0.227 (0.343)
108.0	0.381	8.91 (8.91)	0.211 (0.301)
147.4	0.402	7.91 (7.93)	0.202 (0.270)

Table 14: *AC-SSP* ocean-temperature (*thetao*) performance by depth level, SSP5-3.4-over, last decade of the rollout. SpStd/IAV given as model (IPSL).

[16] Beusch L, Gudmundsson L, Seneviratne SI. Emulating Earth System Model Temperatures with MESMER: From Global Mean Temperature Trajectories to Grid-Point-Level Realizations on Land. *Earth System Dynamics*. 2020 Feb;11(1):139-59.

[17] Leach NJ, Jenkins S, Nicholls Z, Smith CJ, Lynch J, Cain M, et al. FaIRv2.0.0: A Generalized Impulse Response Model for Climate Uncertainty and Future Scenario Exploration. *Geoscientific Model Development*. 2021 May;14(5):3007-36.

[18] Meinshausen M, Raper SCB, Wigley TML. Emulating Coupled Atmosphere-Ocean and Carbon Cycle Models with a Simpler Model, MAGICC6 – Part 1: Model Description and Calibration. *Atmospheric Chemistry and Physics*. 2011 Feb;11(4):1417-56.

[19] Mengis N, Keller DP, MacDougall AH, Eby M, Wright N, Meissner KJ, et al. Evaluation of the University of Victoria Earth System Climate Model Version 2.10 (UVic ESCM 2.10). *Geoscientific Model Development*. 2020 Sep;13(9):4183-204.

[20] Wigley T. Developing Climate Scenarios From Equilibrium GCM Results.

[21] Womack CB, Giani P, Eastham SD, Selin NE. Rapid Emulation of Spatially Resolved Temperature Response to Effective Radiative Forcing. *Journal of Advances in Modeling Earth Systems*. 2025;17(1):e2024MS004523.

[22] Tebaldi C, Arblaster JM. Pattern Scaling: Its Strengths and Limitations, and an Update on the Latest Model Simulations. *Climatic Change*. 2014 Feb;122(3):459-71.

[23] Castruccio S, McInerney DJ, Stein ML, Crouch FL, Jacob RL, Moyer EJ. Statistical Emulation of Climate Model Projections Based on Precomputed GCM Runs. *Journal of Climate*. 2014 Mar;27(5):1829-44.

[24] Acosta MC, Palomas S, Paronuzzi Ticco SV, Utrera G, Biercamp J, Bretonniere PA, et al. The Computational and Energy Cost of Simulation and Storage for Climate Science: Lessons from CMIP6. *Geoscientific Model Development*. 2024 Apr;17(8):3081-98.

[25] Gelbrecht M, White A, Bathiany S, Boers N. Differentiable Programming for Earth System Modeling. *Geoscientific Model Development*. 2023 Jun;16(11):3123-35.

[26] Klocke D, Frauen C, Engels JF, Alexeev D, Redler R, Schnur R, et al. Computing the Full Earth System at 1 Km Resolution. *arXiv*; 2025.

[27] Lapillonne X, Hupp D, Gessler F, Walser A, Pauling A, Lauber A, et al. Operational Numerical Weather Prediction with ICON on GPUs (Version 2024.10). *Geoscientific Model Development*. 2026 Jan;19(2):755-72.

[28] Alet F, Price I, El-Kadi A, Masters D, Markou S, Andersson TR, et al. Skillful Joint Probabilistic Weather Forecasting from Marginals. *arXiv*; 2025.

[29] Price I, Sanchez-Gonzalez A, Alet F, Ewalds T, El-Kadi A, Stott J, et al. GenCast: Diffusion-based Ensemble Forecasting for Medium-Range Weather. *arXiv*; 2023.

- [30] Cresswell-Clay N, Liu B, Durran DR, Liu Z, Espinosa ZI, Moreno RA, et al. A Deep Learning Earth System Model for Efficient Simulation of the Observed Climate. *AGU Advances*. 2025;6(4):e2025AV001706.
- [31] Guan H, Arcomano T, Chattopadhyay A, Maulik R. LUCIE: A Lightweight Uncoupled Climate Emulator With Long-Term Stability and Physical Consistency. *Journal of Advances in Modeling Earth Systems*. 2025;17(11):e2025MS005152.
- [32] Cachay SR, Zhao B, Joren H, Yu R. DYffusion: A Dynamics-informed Diffusion Model for Spatiotemporal Forecasting. *arXiv*; 2023.
- [33] Brenowitz ND, Ge T, Subramaniam A, Gupta A, Hall DM, Mardani M, et al. Climate in a Bottle: Towards a Generative Foundation Model for the Kilometer-Scale Global Atmosphere. *arXiv*; 2025.
- [34] Wang C, Pritchard MS, Brenowitz N, Cohen Y, Bonev B, Kurth T, et al. Coupled Ocean-Atmosphere Dynamics in a Machine Learning Earth System Model. *arXiv*; 2024.
- [35] Duncan JPC, Wu E, Dheeshjith S, Subel A, Arcomano T, Clark SK, et al. SamudrACE: Fast and Accurate Coupled Climate Modeling with 3D Ocean and Atmosphere Emulators. *arXiv*; 2025.
- [36] Dheeshjith S, Subel A, Adcroft A, Busecke J, Fernandez-Granda C, Gupta S, et al. Samudra: An AI Global Ocean Emulator for Climate. *arXiv*; 2025.
- [37] Nguyen T, Brandstetter J, Kapoor A, Gupta JK, Grover A. ClimaX: A Foundation Model for Weather and Climate. *arXiv*. 2023 Jan.
- [38] Watson-Parris D, Rao Y, Oliv   D, Seland  , Nowack P, Camps-Valls G, et al. ClimateBench v1.0: A Benchmark for Data-Driven Climate Projections. *Journal of Advances in Modeling Earth Systems*. 2022 Oct;14(10).
- [39] Kaltenborn J, Lange CEE, Ramesh V, Brouillard P, Gurwicz Y, Nagda C, et al. ClimateSet: A Large-Scale Climate Model Dataset for Machine Learning. *arXiv*; 2023.
- [40] L  tjens B, Ferrari R, Watson-Parris D, Selin N. The Impact of Internal Variability on Benchmarking Deep Learning Climate Emulators. *arXiv*; 2025.
- [41] Clark SK, Watt-Meyer O, Kwa A, McGibbon J, Henn B, Perkins WA, et al. ACE2-SOM: Coupling an ML Atmospheric Emulator to a Slab Ocean and Learning the Sensitivity of Climate to Changed CO₂. *Journal of Geophysical Research: Machine Learning and Computation*. 2025;2(4):e2024JH000575.
- [42] Kochkov D, Yuval J, Langmore I, Norgaard P, Smith J, Mooers G, et al. Neural General Circulation Models for Weather and Climate. *arXiv*; 2024.
- [43] Dewey M, Hansson HC, Watson-Parris D, Samset BH, Wilcox LJ, Lewinschal A, et al. AeroGP: Machine Learning How Aerosols Impact Regional Climate. *Journal of Geophysical Research: Machine Learning and Computation*. 2025;2(4):e2025JH000741.
- [44] Immorlano F, Tavares E, Draxler F, Smyth P, Gentine P, Mandt S. Technical Report: Towards Unified Diffusion Models for Multi-Model Climate Emulation at Scale. *arXiv*; 2025.
- [45] Mansfield LA, Nowack PJ, Kasoar M, Everitt RG, Collins WJ, Voulgarakis A. Predicting Global Patterns of Long-Term Climate Change from Short-Term Simulations Using Machine Learning. *npj Climate and Atmospheric Science*. 2020 Nov;3(1):44.
- [46] Meinshausen M, Vogel E, Nauels A, Lorbacher K, Meinshausen N, Etheridge DM, et al. Historical Greenhouse Gas Concentrations for Climate Modelling (CMIP6). *Geoscientific Model Development*. 2017 May;10(5):2057-116.
- [47] Canadell JG, Monteiro PM, Costa MH, Syampungani S, Zaehle S, Zickfeld Canada K, et al. 2021: Global Carbon and Other Biogeochemical Cycles and Feedbacks. In *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*; 2021.

- [48] Niemeier U, Timmreck C. What Is the Limit of Climate Engineering by Stratospheric Injection of SO₂? *Atmospheric Chemistry and Physics*. 2015 Aug;15(16):9129-41.
- [49] Feichter J, Roeckner E, Lohmann U, Liepert B. Nonlinear Aspects of the Climate Response to Greenhouse Gas and Aerosol Forcing. *Journal of Climate*. 2004 Jun;17(12):2384-98.
- [50] Aerosol in the Climate System. In: *Aerosols and Climate*. Elsevier; 2022. p. 9-52.
- [51] Lurton T, Balkanski Y, Bastrikov V, Bekki S, Bopp L, Braconnot P, et al. Implementation of the CMIP6 Forcing Data in the IPSL-CM6A-LR Model. *Journal of Advances in Modeling Earth Systems*. 2020;12(4):e2019MS001940.
- [52] Hauglustaine DA, Hourdin F, Jourdain L, Filiberti MA, Walters S, Lamarque JF, et al. Interactive Chemistry in the Laboratoire de Météorologie Dynamique General Circulation Model: Description and Background Tropospheric Chemistry Evaluation. *Journal of Geophysical Research: Atmospheres*. 2004;109(D4).
- [53] Pyle J, Shepherd T, Bodeker G, Gruzdev A, Müller R, Muthama NJ, et al. Ozone and Climate: A Review of Interconnections.
- [54] Laboratory (CSL) NCS. Scientific Assessment of Ozone Depletion 2018: Executive Summary; 2018. <https://csl.noaa.gov/assessments/ozone/2018/executivesummary/#>.
- [55] Morgenstern O, Stone KA, Schofield R, Akiyoshi H, Yamashita Y, Kinnison DE, et al. Ozone Sensitivity to Varying Greenhouse Gases and Ozone-Depleting Substances in CCMI-1 Simulations. *Atmospheric Chemistry and Physics*. 2018 Jan;18(2):1091-114.
- [56] Nowack PJ, Abraham NL, Braesicke P, Pyle JA. The Impact of Stratospheric Ozone Feedbacks on Climate Sensitivity Estimates. *Journal of Geophysical Research: Atmospheres*. 2018;123(9):4630-41.
- [57] Ma Y, Abraham NL, Versick S, Ruhnke R, Schneider A, Niemeier U, et al. Mloz: A Highly Efficient Machine Learning-Based Ozone Parameterization for Climate Sensitivity Simulations. *Journal of Advances in Modeling Earth Systems*. 2026;18(4):e2025MS005459.
- [58] Son SW, Gerber EP, Perlwitz J, Polvani LM, Gillett NP, Seo KH, et al. Impact of Stratospheric Ozone on Southern Hemisphere Circulation Change: A Multimodel Assessment. *Journal of Geophysical Research: Atmospheres*. 2010;115(D3).
- [59] Wang J, Chiodo G, Ayarzagüena B, Ball WT, Diallo M, Hassler B, et al. Exploring Ozone–Climate Interactions in Idealized CMIP6 DECK Experiments. *Atmospheric Chemistry and Physics*. 2025 Dec;25(23):17819-44.
- [60] Dickau M, Zickfeld K, Matthews HD. Irreversible Climate Changes Driven by Degree-Years of Temperature Overshoot. *Research Square*; 2025.
- [61] Hourdin F, Rio C, Grandpeix JY, Madeleine JB, Cheruy F, Rochetin N, et al. LMDZ6A: The Atmospheric Component of the IPSL Climate Model With Improved and Better Tuned Physics. *Journal of Advances in Modeling Earth Systems*. 2020;12(7):e2019MS001892.
- [62] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. *arXiv*; 2021.
- [63] Bi K, Xie L, Zhang H, Chen X, Gu X, Tian Q. Pangu-Weather: A 3D High-Resolution Model for Fast and Accurate Global Weather Forecast. *arXiv*; 2022.
- [64] AIFS-CRPS: Ensemble Forecasting Using a Model Trained with a Loss Function Based on the Continuous Ranked Probability Score | *Npj Artificial Intelligence*; 2026. <https://www.nature.com/articles/s44387-026-00073-7>.
- [65] Lang S, Leutbecher M, Hatfield S. On the Sensitivity of Machine-Learned Probabilistic Weather Forecast Models to Scale-Aware Scoring Rules. *arXiv*; 2026.
- [66] Park T, Liu MY, Wang TC, Zhu JY. Semantic Image Synthesis with Spatially-Adaptive Normalization. *arXiv*; 2019.

- [67] Chen M, Tan X, Li B, Liu Y, Qin T, Zhao S, et al. AdaSpeech: Adaptive Text to Speech for Custom Voice. *arXiv*; 2021.
- [68] Gneiting T, Raftery AE. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*. 2007 Mar;102(477):359-78.
- [69] Lang S, Alexe M, Clare MCA, Roberts C, Adewoyin R, Ben Bouallègue Z, et al. AIFS-CRPS: Ensemble Forecasting Using a Model Trained with a Loss Function Based on the Continuous Ranked Probability Score. *npj Artificial Intelligence*. 2026 Feb;2(1):18.
- [70] Howard LJ, Barnes EA. Do AI Forecast Ensembles Sample the Correct Conditional Distribution?. *arXiv*; 2026.
- [71] Scheuerer M, Hamill TM. Variogram-Based Proper Scoring Rules for Probabilistic Forecasts of Multivariate Quantities. *Monthly Weather Review*. 2015 Apr;143(4):1321-34.
- [72] Holton JR, Hakim GJ. Basic Conservation Laws. In: *An Introduction to Dynamic Meteorology*. Elsevier; 2013. p. 31-66.
- [73] Trenberth KE. An Imperative for Climate Change Planning: Tracking Earth's Global Energy. *Current Opinion in Environmental Sustainability*. 2009 Oct;1(1):19-27.
- [74] von Schuckmann K, Cheng L, Palmer MD, Hansen J, Tassone C, Aich V, et al. Heat Stored in the Earth System: Where Does the Energy Go? *Earth System Science Data*. 2020 Sep;12(3):2013-41.
- [75] Beusch L, Nicholls Z, Gudmundsson L, Hauser M, Meinshausen M, Seneviratne SI. From Emission Scenarios to Spatially Resolved Projections with a Chain of Computationally Efficient Emulators: Coupling of MAGICC (v7.5.1) and MESMER (v0.8.3). *Geoscientific Model Development*. 2022 Mar;15(5):2085-103.
- [76] Rose BEJ, Rayborn L. The Effects of Ocean Heat Uptake on Transient Climate Sensitivity. *Current Climate Change Reports*. 2016 Dec;2(4):190-201.
- [77] Deutloff J, Held H, Lenton TM. High Probability of Triggering Climate Tipping Points under Current Policies Modestly Amplified by Amazon Dieback and Permafrost Thaw. *Earth System Dynamics*. 2025 Apr;16(2):565-83.
- [78] Previdi M, Smith KL, Polvani LM. Arctic Amplification of Climate Change: A Review of Underlying Mechanisms. *Environmental Research Letters*. 2021 Sep;16(9):093003.
- [79] Swart NC, Cole JNS, Kharin VV, Lazare M, Scinocca JF, Gillett NP, et al. The Canadian Earth System Model Version 5 (CanESM5.0.3). *Geoscientific Model Development*. 2019 Nov;12(11):4823-73.
- [80] Clark SK, Arcomano T, Duncan JPC, Henn B, Kwa A, McGibbon J, et al. Disentangling the Effects of Sea Surface Temperature and CO₂ in Global Machine Learned Weather-Climate Emulators. *arXiv*; 2026.
- [81] Zanchettin D, Khodri M, Timmreck C, Toohey M, Schmidt A, Gerber EP, et al. The Model Intercomparison Project on the Climatic Response to Volcanic Forcing (VolMIP): Experimental Design and Forcing Input Data for CMIP6. *Geoscientific Model Development*. 2016 Aug;9(8):2701-19.
- [82] Timmreck C, Mann GW, Aquila V, Hommel R, Lee LA, Schmidt A, et al. The Interactive Stratospheric Aerosol Model Intercomparison Project (ISA-MIP): Motivation and Experimental Design. *Geoscientific Model Development*. 2018 Jul;11(7):2581-608.
- [83] Hickman S, Trajkovic I, Kaltenborn J, Pelletier F, Archibald A, Gurwicz Y, et al. Causal Climate Emulation with Bayesian Filtering. *arXiv*; 2025.
- [84] Su J, Lu Y, Pan S, Murtadha A, Wen B, Liu Y. RoFormer: Enhanced Transformer with Rotary Position Embedding. *arXiv*; 2023.
- [85] Stock J, Arcomano T, Kotamarthi R. Swift: An Autoregressive Consistency Model for Efficient Weather Forecasting. *arXiv*; 2025.

- [86] Levitus S, Antonov JI, Boyer TP, Baranova OK, Garcia HE, Locarnini RA, et al. World Ocean Heat Content and Thermosteric Sea Level Change (0–2000 m), 1955–2010. *Geophysical Research Letters*. 2012;39(10).
- [87] Griffies SM, Danabasoglu G, Durack PJ, Adcroft AJ, Balaji V, Böning CW, et al. OMIP Contribution to CMIP6: Experimental and Diagnostic Protocol for the Physical Component of the Ocean Model Intercomparison Project. *Geoscientific Model Development*. 2016 Sep;9(9):3231-96.
- [88] Myhre G, Highwood EJ, Shine KP, Stordal F. New Estimates of Radiative Forcing Due to Well Mixed Greenhouse Gases. *Geophysical Research Letters*. 1998;25(14):2715-8.
- [89] Etminan M, Myhre G, Highwood EJ, Shine KP. Radiative Forcing of Carbon Dioxide, Methane, and Nitrous Oxide: A Significant Revision of the Methane Radiative Forcing. *Geophysical Research Letters*. 2016;43(24):12, 614-2, 623.
- [90] Gregory JM, Forster PM. Transient Climate Response Estimated from Radiative Forcing and Observed Temperature Change. *Journal of Geophysical Research: Atmospheres*. 2008;113(D23).