

A review of spatially explicit climate emulators as agile tools for decision-support

Sarah Schoengart

sarah.schoengart@env.ethz.ch

ETH Zurich <https://orcid.org/0009-0002-2030-9139>

Lukas Gudmunsson

ETH Zurich

Christopher B. Womack

ETH Zurich

Carl-Friedrich Schleussner

International Institute for Applied Systems Analysis

Sonia I. Seneviratne

ETH Zurich

Systematic Review

Keywords: earth system model, emulator, climate information

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11205768/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

A review of spatially explicit climate emulators as agile tools for decision-support

**Sarah Schöngart^{a,d}, Lukas Gudmundsson^a, Christopher B. Womack^{b,e},
Carl-Friedrich Schleussner^c, and Sonia I. Seneviratne^a**

^aInstitute for Atmospheric and Climate Science, ETH Zürich, Zürich, Switzerland

^bDepartment of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA, USA

^cInternational Institute for Applied Systems Analysis, Laxenburg, Austria

^dIRIThesis, Humboldt-Universität zu Berlin, Friedrichstrasse 191, 10117 Berlin, Germany

^eCenter for Sustainability Science and Strategy, Massachusetts Institute of Technology, Cambridge, MA, USA

Corresponding author:

Sarah Schöngart¹

Email address: sarah.schoengart@env.ethz.ch

ABSTRACT

Spatially explicit emulators have emerged as agile tools for generating climate and weather information at the speed and scale required for decision-making. However, their current applicability for decision support is limited by a fragmented landscape and an opaque, quickly shifting boundary between climate and weather emulators. This shift is driven by the emergence of large-scale foundation models and by architectural advances that allow short-term forecasting tools to increasingly target climate timescales. This convergence calls for a systematic evaluation of their intersection and relative skill. Through a review of 76 studies, we categorise spatially explicit emulators by their skill across weather and climate timescales, and organise frameworks by inputs, outputs, methods, and evaluation practices. We identify three key patterns. First, the field is siloed: weather emulators form a coherent, deep learning-driven cluster, while climate emulators remain diverse and weakly connected to these advances. Despite claims of cross-timescale applicability, weather and hybrid emulators lack robust validation under long-term anthropogenic forcing. Second, emulator development and application remain weakly aligned. Climate emulators prioritise scenario agility over output flexibility, limiting application-specific use, while restricted accessibility results in limited uptake, with most applications confined to their developers. Third, many state-of-the-art weather emulators rely on specialised infrastructure and are often industry-led, raising concerns about computational equity and “agility for whom.” Realising genuine agility across the weather–climate spectrum and its utility for decision support requires clearer task definitions, stronger cross-domain integration, and deliberate attention to equitable access.

Keywords: Earth System Model; Emulator; Decision-relevant climate information

From global to regional scales, decisions today are fundamentally influenced by realised and anticipated changes in the climate system^[1,2]. This is increasingly evident in fields like end-to-end climate attribution, where linking extreme events to specific emission sources now informs corporate risk assessment, litigation, and climate financing mechanisms^[3–5]. Supporting these and other emerging applications requires an agile climate tool landscape, one capable of flexible scenario design and thorough risk-space sampling using large ensembles^[6–8]. Yet, the traditional computational regime, centred on Earth System Models (ESMs) and physics-based models that prioritise mechanistic realism, remains prohibitively costly. Individual simulations often require weeks to months and produce terabytes of output^[9,10], effectively limiting their utility in the very data-intensive or time-sensitive contexts these new applications demand^[11].

This computational barrier has spurred the development of a new data-driven paradigm in climate science to complement traditional tools^[12], leveraging approaches that interact with and build upon empirical observation, theory, and computational modelling (Fig. 1, panel a). Among these emerging methods are

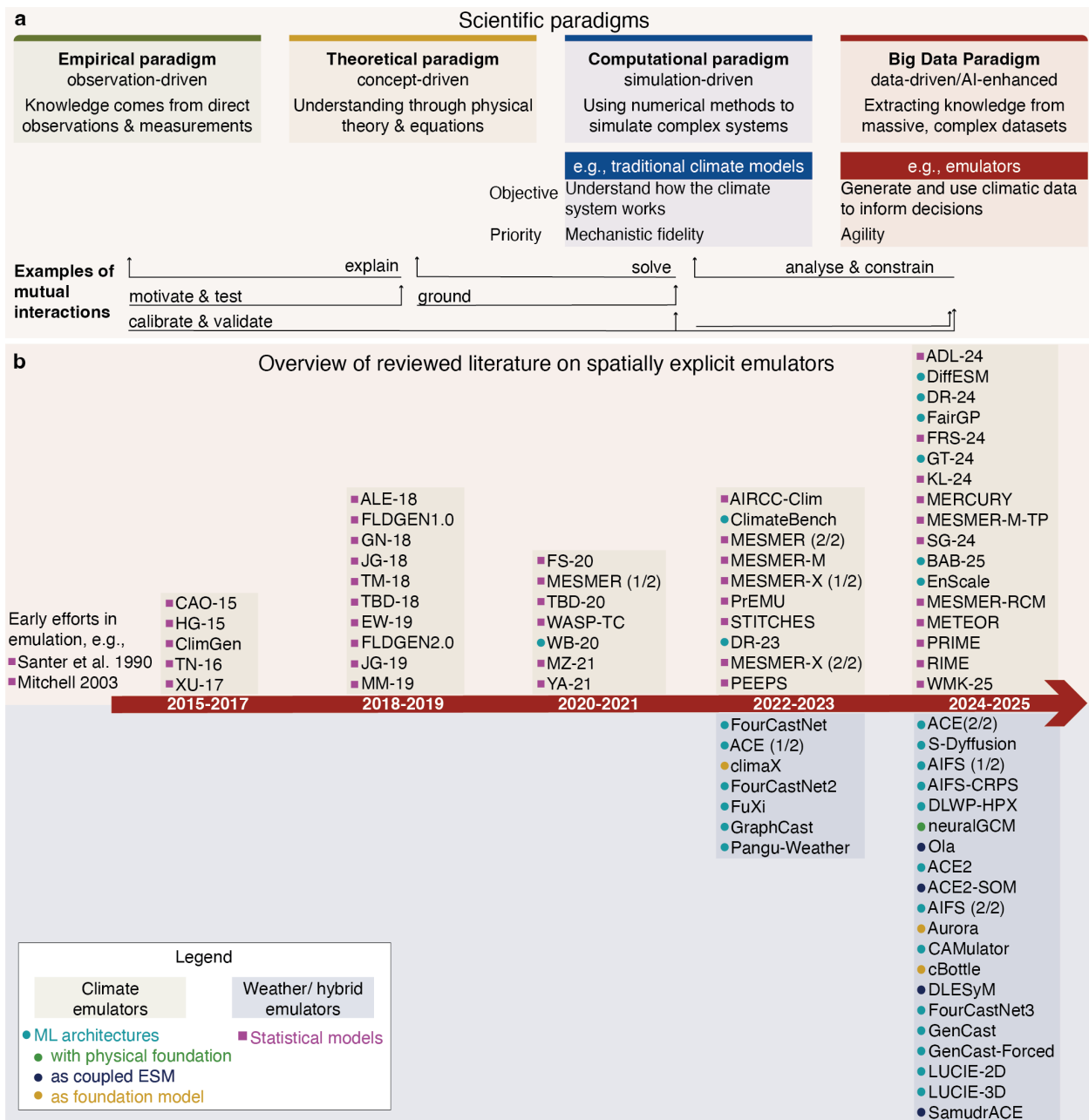


Figure 1. Defining the scope of the review. **a**, The four scientific paradigms—empirical observation (green), theoretical insight (yellow), computational modeling (blue), and data-driven approaches (red)—function as complementary components of scientific inquiry, with interactions indicated by arrows. For example, observations are used to calibrate and validate models, while models guide and extend observational understanding. Traditional climate models fall within the computational paradigm, emulators fall within the big data paradigm. **b**, The review includes papers on spatially explicit climate and weather emulators published over the past decade (2015-2025) and spans the weather to climate domain. Each emulation approach is indicated by its name or short identifier (see Table 1) and positioned according to its publication year (within each bin, papers are grouped by year and ordered alphabetically with publications from the first year in the bin appearing on top). The color coding indicates the emulation category (weather/hybrid or climate) and the underlying methodology (machine learning or statistical methods).

climate emulators—statistical or machine learning (ML) models trained to approximate selected outputs of ESMs. These emulators allow for the rapid generation of climate projections, often in seconds to minutes, thereby enabling fast scenario exploration, uncertainty quantification, and regional impact assessments^[13]. They operate by trading mechanistic realism for computational efficiency, making them particularly suitable for decision contexts where quantifying probabilities and assessing risk take precedence over analysing and understanding the underlying physical drivers.

The agility offered by emulators can support scientific assessments such as those of the Intergovernmental Panel on Climate Change (IPCC), as demonstrated by the use of simple climate modelling approaches to assess global mean quantities^[14,15]. The rise of spatially explicit tools extends these approaches, and could allow to represent scenario-to-climate linkages at scales relevant for regional and local decision-making^[16].

Beyond long-term climate applications, the rapid advancement of deep learning techniques is dissolving the conventional divide between models focused on short-term weather forecasting and those designed for long-term climate projection. State-of-the-art weather emulators—originally designed for short-range forecasting (up to 15 days)—are now incorporating external forcing, such as sea surface temperature (SST) or CO₂ concentrations, enabling stability over seasonal to interannual timescales^[17,18] and coupled weather-ocean emulators are emerging to build fully coupled ESM-style ML models^[19,20]. Meanwhile, dedicated climate emulators continue to develop independently, using statistical or ML methods to emulate long-term responses to anthropogenic forcing^[11,21]. While the boundaries between weather and climate emulation seem to become increasingly blurred, these threads evolve largely in parallel, with limited shared terminology, inconsistent evaluation standards, and little cross-community integration.

Here, we clarify and connect these diverse development strands of the emulator literature, with a focus on how they impact climate emulation. Through a literature analysis of spatially explicit emulators developed over the past decade (Fig. 1, panel b), we identify the conceptual and methodological continuum connecting weather to climate emulators. We start by formally defining climate emulation, and then organise the landscape of existing approaches into the three subspaces of weather, hybrid weather–climate, and climate emulation. We then examine applications of emulated data, highlighting opportunities, current limitations, and key development priorities, before concluding with a discussion of our findings.

Box 1: Terminology

Weather models **forecast**, predicting the specific evolution of the atmosphere over hours to days. In contrast, climate models **project**, focusing on the statistical response of the climate system to prescribed external forcing over decades to centuries. The term simulation describes any model-generated trajectory in both contexts.

Across these tasks, models rely on three types of variables. **Forcing variables** act purely as inputs, such as greenhouse gas (GHG) emissions or solar irradiance. **Prognostic variables** serve as both inputs and outputs; the model advances these variables forward in time, and the resulting state forms the input for the next timestep (e.g., temperature, pressure, wind velocity, and humidity). **Diagnostic variables** are derived quantities that describe aspects of the model state but appear only as outputs (e.g., precipitation, cloud cover).

Since emulators are data-driven models, their development fundamentally depends on the availability of suitable datasets, either reanalysis products that assimilate observations into numerical models (e.g., ERA5^[22], MERRA-2^[23]) or direct outputs from physics-based models. These datasets are typically partitioned into **training** data, used to fit (or calibrate) model parameters; **validation** data, used to tune hyper-parameters that govern the learning process or model structure—and to monitor overfitting; and **test** data, which are held out for the final evaluation of performance on unseen samples. This workflow enables transparent model assessment and guards against overly optimistic performance estimates.

Table 1. Overview of reviewed emulator literature. Papers are organised by topic: weather (first column), hybrid weather–climate (second column), and climate (remaining columns). Hybrid weather–climate approaches are functionally similar to weather emulators, but typically include external forcing and remain stable over extended or historic periods. Climate emulators are further grouped by emulator objective. Emulators for scenario exploration target long-term responses, scenario specific emulators act as tools to generate larger ensembles for individual scenarios or to compress a specific scenario with generative decompression, and extension approaches build on gridded variables to increase either the number of variables or the spatiotemporal resolution. Approaches are organised alphabetically within each column.

Weather emulators	Weather–climate emulators	Climate emulators		
		Scenario exploration	Scenario specific	Extensions
AIFS ^[24,25] , AIFS-CRPS ^[26]	ACE ^[27,28]	ADL-24 ^[29]	EW-19 ^[30]	DiffESM ^[31]
Aurora ^[32]	ACE2 ^[18]	AIRCC-Clim ^[33]	FS-20 ^[34]	DR-23 ^[35]
DLWP-HPX ^[36]	ACE2-SOM ^[19]	ALE-18 ^[37]	GN-18 ^[38]	DR-24 ^[39]
FourCastNet ^[40–42]	CAMulator ^[43]	BAB-25 ^[44]	GT-24 ^[45]	EnScale ^[46]
FuXi ^[47]	cBottle ^[48]	CAO-15 ^[49]	JG-18 ^[50]	MESMER-M ^[51]
GenCast ^[52] , GenCast-forced ^[53]	climaX ^[54]	ClimateBench ^[21]	JG-19 ^[55]	MESMER-M-TP ^[56]
GraphCast ^[57]	DLESyM ^[20]	ClimGen ^[58]	MM-19 ^[59]	MESMER-RCM ^[60]
Pangu-Weather ^[61]	LUCIE-2D ^[62]	FairGP ^[63]	MZ-2021 ^[64]	PrEmu ^[65]
	LUCIE-3D ^[66]	FLDGEN1.0 ^[67]	WB-20 ^[68]	
	neuralGCM ^[17]	FLDGEN2.0 ^[69]		
	Ola ^[70]	FRS-24 ^[71]		
	SamudrACE ^[72]	HG-15 ^[73]		
	S-Dyffusion ^[74]	KL-24 ^[75]		
		MERCURY ^[76]		
		MESMER ^[11,77]		
		MESMER-X ^[78,79]		
		METEOR ^[80]		
		PEEPS ^[81]		
		PRIME ^[82]		
		RIME ^[83]		
		SG-24 ^[84]		
		STITCHES ^[85]		
		TBD-18 ^[86] , TBD-20 ^[87]		
		TN-16 ^[88]		
		TM-18 ^[89]		
		WASP-TC ^[90]		
		WMK-25 ^[91]		
		XU-17 ^[92]		
		YA-21 ^[93]		

FROM NUMERICAL MODELS TO EMULATORS

This section establishes the conceptual foundation of climate emulation by first introducing the traditional physics-based modelling paradigm and then reinterpreting it through a probabilistic lens, which naturally motivates the development of data-driven emulators.

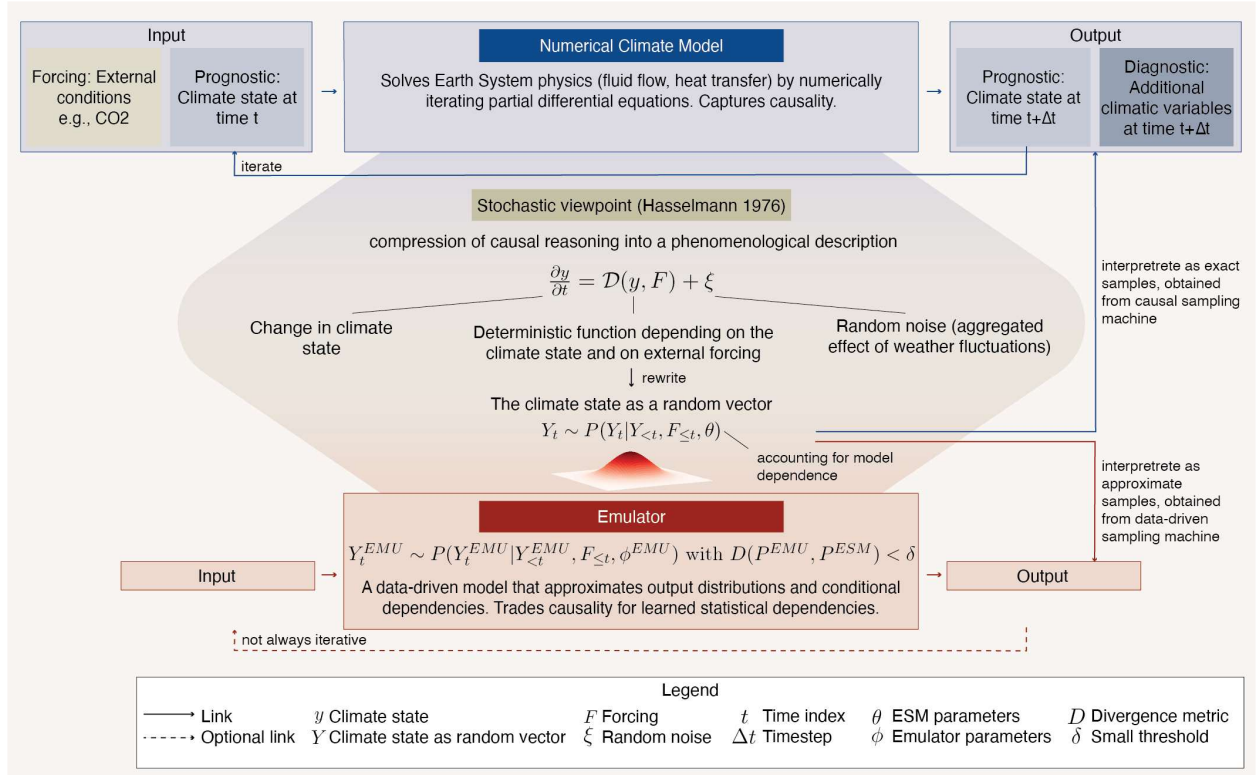


Figure 2. Conceptual framework of climate emulators. Numerical climate models, e.g., ESMs, take external forcing and an initial climate state and iteratively advance the state through process-based numerical descriptions of the Earth system. The output includes the advanced initial state (prognostic, serves as input and output) and additional diagnostic (output-only) variables. Hasselmann^[94] interpret the model-induced relationship between input and output stochastically through stochastic differential equations, allowing it to describe the climate state as a multivariate probability distribution conditional on past states, the history of external forcing, and model parameters. Emulators learn these distributions from ESM samples and generate new ones for unseen forcings or initial states.

The traditional numerical modelling paradigm Climate modelling traditionally relies on ESMs which simulate the coupled physical, chemical and biological process of the planet^[95]. These physics-based models use numerical implementations of physical laws—such as fluid dynamics—to evolve the system state forward in time (Fig. 2). At the scale of climate, the system’s evolution is primarily governed by boundary conditions, for example, prescribed external forcing such as GHGs and aerosols, which act over long timescales. In contrast, weather prediction focuses on much shorter timescales, where the primary input is the initial, highly accurate description of the current system state and changes in external forcing are negligible. Thus, while weather prediction is fundamentally an initial condition problem, climate projections are fundamentally a boundary condition problem (Fig. 3, panel a)^[96]. In either case, the model simulates the system through the implemented equations and outputs the next timestep. Although the underlying equations are deterministic in theory, imperfect initial conditions, structural model uncertainty and numerical approximations limit predictability across all timescales.

Interpreting numerical models as stochastic sampling machines Numerical models, such as ESMs, are process-based, using causal reasoning to evolve the climate state forward in time (Fig. 2). Hasselmann^[94] provided a crucial abstraction of this view by separating slow, deterministic dynamics from fast, apparently random fluctuations. They formulated climate evolution as a stochastic differential equation:

$$\frac{\partial y}{\partial t} = \mathcal{D}(y, F) + \xi, \quad (1)$$

which interprets the change in climatic state, y , as the result of a deterministic evolution, $\mathcal{D}$, that depends on external forcing, F , and the climatic state itself, plus a random term, ξ , which represents aggregated fast effects, including weather and interannual variability. This formulation shifts the perspective from explaining why the system changes to describing how it changes, transforming a causal, process-based view into a phenomenological one (Fig. 2).

Climate model output exhibits path dependence and memory, driven by slowly adjusting components like the ocean. The output also captures internal variability, arising from apparently random fluctuations and coherent patterns such as El Niño^[97,98]. The abstraction from Hasselmann^[94], allows us to interpret the climate state as a random vector whose probability distribution depends on its past and on external forcing, effectively encapsulating the system’s physics within conditional dependencies:

$$\mathbf{Y}_t \sim P(\mathbf{Y}_t \mid \mathbf{Y}_{<t}, F_{\leq t}, \theta), \quad (2)$$

where θ denotes model-specific structural parameters (see Womack et al.^[91] for an explicit linkage from Eq. 1 to Eq. 2). In this probabilistic formulation, the temporal evolution of the climate is embedded in conditional dependencies rather than explicit dynamical equations.

Using this abstraction, ESMs can be viewed as sampling machines that explore the conditional distribution in Eq. 2. Coordinated modelling efforts, such as the Coupled Model Intercomparison Project (CMIP), explicitly account for these conditional dependencies through distinct ensemble designs:

- **Initial-condition ensembles** sample internal variability by systematically varying the initial climate state $\mathbf{Y}_{<t}^{(i)}$ (achieved by changing initial conditions $\mathbf{Y}_{t_0}^{(i)}$) where i indexes ensemble members.
- **Multi-model ensembles** sample structural uncertainty by employing different climate models, each defined by a distinct parameter set $\theta^{(j)}$, with j indexing models.
- **Scenario ensembles** sample forcing (scenario) uncertainty^[99] by fixing the model and initial conditions while varying external forcing $F_{\leq t}^{(k)}$, with k indexing scenarios.

Defining the emulation task The reinterpretation of ESMs as stochastic sampling devices naturally motivates the development of emulators. An emulator is a statistical or ML model trained to reproduce the distributional properties of ESM output. In essence, it is an alternative, data-driven model that samples from the distribution

$$\mathbf{Y}_t^{\text{EMU}} \sim P(\mathbf{Y}_t^{\text{EMU}} \mid \mathbf{Y}_{<t}^{\text{EMU}}, F_{\leq t}, \phi^{\text{EMU}}), \quad (3)$$

such that the probability distribution sampled by the emulator closely approximates that of the ESM, as measured by a divergence metric D and a tolerance threshold δ :

$$D(P^{\text{EMU}}, P^{\text{ESM}}) < \delta. \quad (4)$$

We denote model parameters from physics-based models using θ and from emulators using ϕ . We also use the superscripts (ESM/EMU) whenever necessary to make the distinction between models clear. Common divergence metrics include the root mean squared error (RMSE)^[42], energy distance e.g.,^[46], and continuous ranked probability score (CRPS) e.g.,^[21,42,78,79].

Emulators are calibrated on existing ESM output to provide a fast and flexible means of generating new climate samples, effectively extending the concept of ESMs as stochastic samplers into a data-driven form. The main challenge lies in approximating a complex, multivariate probability distribution from a finite set of ESM samples. Optimizing the full joint log-likelihood of climate data over long temporal sequences is typically intractable. While simpler emulators may be optimized solely on low-order moments (e.g., the MSE of the mean response field^[81]), modern generative models like diffusion models are optimized to match the conditional probability distribution (e.g., $p(y_{t+1}|y_t)$) via a proxy for the conditional log-likelihood (e.g., a denoising loss)^[74]. Consequently, high-order statistics and long-term correlations (such as those emerging from extended autoregression) are usually evaluated and validated *a posteriori*.

Differences between emulators and reduced complexity climate models with global focus The climate modelling landscape also includes reduced-complexity models, for example with global focus such as Simple Climate Models (SCMs). The modelling philosophy behind SCMs differs fundamentally from that of emulators: SCMs are still process-based, approximating physical laws through simplification, whereas emulators are mainly data-driven, sometimes with physical constraints, and approximate the probability distribution of the output itself. SCMs are driven by prescribed forcings (e.g., GHG emissions, aerosols, or solar irradiance) and produce outputs including GHG concentrations, radiative forcing, and aggregated global-scale variables such as global mean temperature (GMT), ocean heat uptake, or sea-level rise^[100,101]. Their efficiency stems from coarse spatial resolution and the use of global or hemispheric scales with annual mean values, making them orders of magnitude faster than ESMs^[102].

For emulating the climate, spatial explicit emulators are often coupled to SCMs output variables such as GMT or radiative forcing (see Fig. 3).

MAPPING THE SPACE FROM WEATHER TO CLIMATE EMULATORS

Emulators can be designed for weather, climate, or the space between. Although we are primarily interested in climate emulation, the rise of ML-based approaches has blurred the boundary between these tasks. Many recent tools are explicitly framed as models for both weather and climate^[17,18], and emerging foundation models aim to unify weather and climate representations within a single framework^[32,48,54].

Here, we seek to understand both the differences and the areas of overlap across existing emulators when it comes to their skill in emulating the climate. To this end, we reviewed 76 studies on spatially explicit emulation and grouped them into three categories: weather emulators, which focus on timescales of up to roughly 15 days; hybrid weather–climate emulators, which perform well on weather tasks while maintaining stability over climate timescales (10–1000 years) in idealized or historical settings; and climate emulators, which prioritize fidelity over climate timescales under anthropogenic forcing, particularly for future projection tasks (Fig. 3 and Table 1). We organize the literature within each category with respect to their input–output structure, their training and evaluation strategies, and their underlying methodological choices. Throughout this section, the notation a/b reads “a out of b” (e.g., 8/9 reads “8 out of 9”).

WEATHER EMULATORS

Weather emulators are highly advanced deep-learning models that excel at short- to medium-range forecasting by learning to advance the atmospheric state autoregressively. As such, they emulate processes rather than outputs (Fig. 3, panel b). Building on the work of Weyn et al.^[103], a suite of ML-based emulators—including FourCastNet^[40–42], Pangu-Weather^[61], GraphCast^[57], FuXi^[47], DLWP-HPX^[36], AIFS^[24–26], GenCast^[52], and Aurora^[32]—has rapidly emerged (see Supplementary Table S1). These models have already demonstrated the ability to match or outperform traditional numerical weather prediction systems in terms of Root Mean Square Error (RMSE), Anomaly Correlation Coefficient (ACC), and probabilistic scores such as the Continuous Ranked Probability Score (CRPS) when verified against ERA5 reanalysis data. Much of this progress has been driven by advances in ML and substantial investment from major industry actors such as NVIDIA, Google, Microsoft, and Huawei. Among the reviewed approaches, FuXi^[47] and AIFS^[24–26] stand out as the only models developed entirely within public research organisations.

Inputs and outputs All of the aforementioned approaches operate autoregressively, learning to advance the weather state forward in time (Fig. 3, panel a). FourCastNet^[40–42] and Pangu-Weather^[61] perform a first-order autoregression, while all other models employ second-order process. Weather emulators generally exclude external forcing, or, when included, represent only fixed components such as topography^[52] or total incoming solar radiation (TISR) required to capture seasonal cycles^[36]. Without non-stationary external forcing, these emulators are suitable only for short- to medium-range forecasting (up to 10–15 days). Incorporating time-varying forcing, such as SST (e.g., Antonio et al.^[53] adapted GenCast for seasonal forecasting), shifts the emulator into the hybrid weather-climate world (Fig. 3 & 4).

Weather emulators produce global, gridded data on three-dimensional meshes (latitude, longitude, vertical levels), focusing on high spatio-temporal resolution. Resolution varies, with Aurora^[32] reaching the highest horizontal resolution ($0.1^\circ \times 0.1^\circ$ latitude-longitude) and Pangu-Weather^[61] offering the highest temporal resolution (1-hourly). 5/8 emulator families operate on $0.25^\circ \times 0.25^\circ$ horizontal grids with 6-hourly resolution (Supplementary Table S1). Outputs universally include temperature and geopotential, and all but one

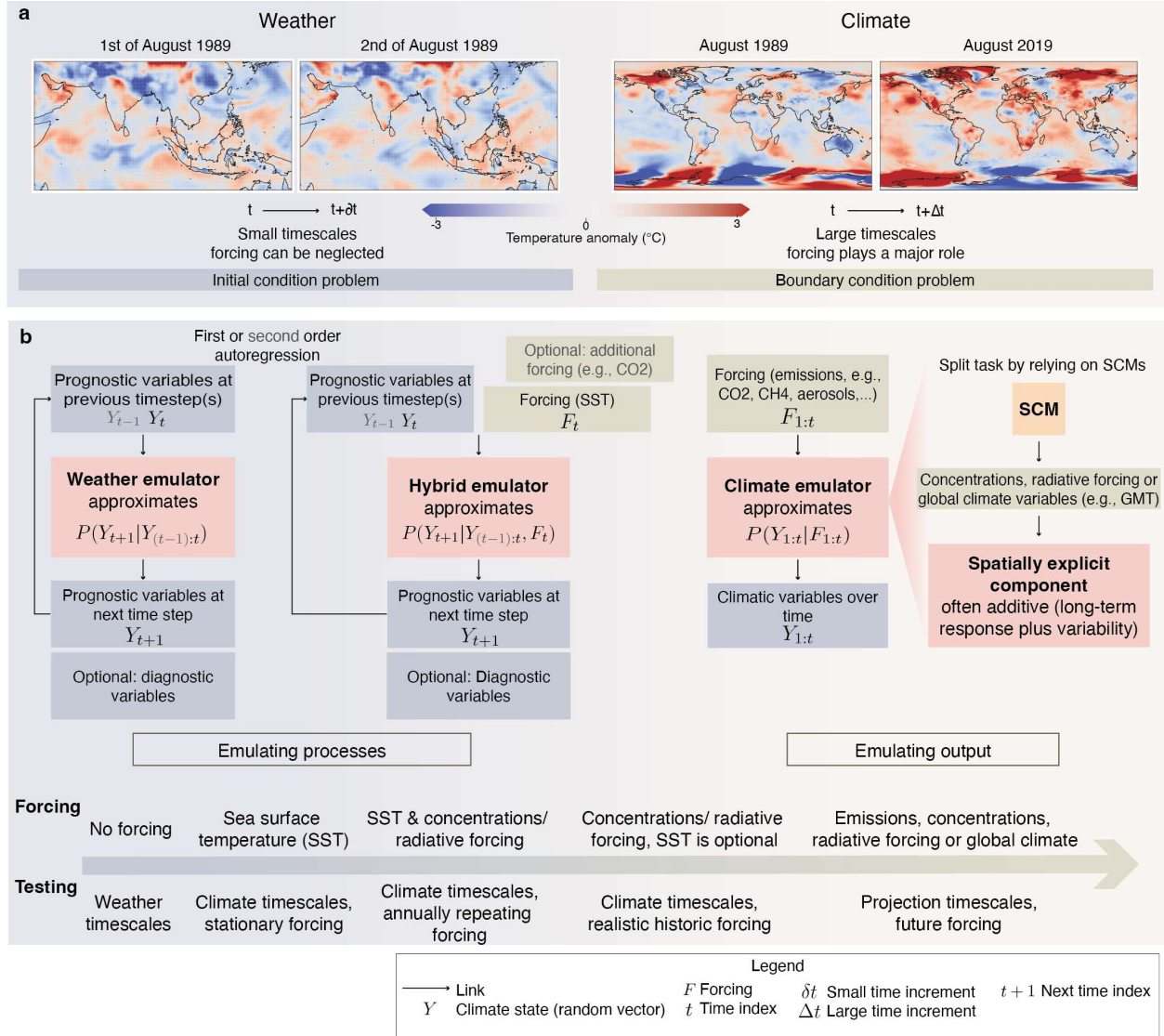


Figure 3. Conceptual design of emulator frameworks. **a**, Weather prediction focuses on short timescales dominated by internal dynamics, while climate emulators focus on long timescales where thermodynamic changes play a major role. Weather is an initial condition problem, whereas climate is primarily driven by boundary conditions. **b**, Conceptual differences among emulator types. Left: Weather emulators use only prognostic variables (the current atmospheric state) and advance them autoregressively to forecast the next state. Middle: Hybrid (AMIP-style) emulators retain this autoregressive structure but incorporate time-varying external forcing, most commonly SST, enabling stable long-term integration. Right: Climate (ESM) emulators rely mainly on external forcing (e.g., emissions, concentrations, or global climate indicators) and typically generate output non-autoregressively, often separating the climate signal into long-term mean response and internal variability. Weather and hybrid emulators are dominated by ML and focus on dynamical processes, whereas ESM emulators are more statistical, focusing on long-term aggregated outputs. The arrow illustrates the methodological trajectory from weather to climate emulators, emphasizing the incorporation of external forcing and adjustments in testing strategies for long-term stability and projection capability.

framework also offer precipitation, wind, humidity, and sea-level pressure. In terms of design philosophy, DLWP-HPX^[36] stands out as the most lightweight approach with the least amount of variables at $1^\circ \times 1^\circ$

latitude-longitude resolution.

Training and evaluation The performance of weather emulators is assessed through hindcasts: the models are initialized with historical conditions and evaluated on their ability to reproduce the weather days ahead. Consequently, training universally relies on the ERA5^[22] reanalysis dataset, with Aurora, designed as a foundation model, being the only one that incorporates diverse data sources into pretraining^[32] (Fig. 4 and Supplementary Table S2). Regardless of the architecture, data splits are defined by calendar year (e.g., FourCastNet^[40] uses 1979–2015 for training, 2016–2017 for validation, and 2018 onward for testing) to ensure models are assessed on temporally unseen data. Notably, several weather emulators—including Pangu-Weather^[61], GraphCast^[57], GenCast^[52], and Aurora^[32]—already outperform the ECMWF IFS HRES^[104], the gold standard in weather modelling, across multiple targets.

Methods Weather emulators are fundamentally deep-learning based and exclusively employ an autoregressive design that mirrors the logic of numerical weather prediction. In this sense, weather models emulate the (stochastic) differential equation that evolves the current weather state into the next, meaning the focus lies on emulating a dynamical process rather than on emulating output distributions. Due to their high resolution and deep architectures, these models require immense computational resources. For example, Pangu-Weather has approximately 256 million parameters, and its training required resources equivalent to over 12,000 V100-equivalent GPU-days (about 64 days on 192 high-end GPUs)^[61] (see Supplementary Section 2). This vast parameter complexity and the associated training time are a key distinction from the computationally lighter climate emulators. However, inference costs are low: Pangu-Weather takes 1.4 s on a single GPU to generate a 24-h forecast, making it more than 10,000-times faster than the IFS numerical weather forecast. The dominant architectures include Adaptive Fourier Neural Operators^[40], Transformers^[61], Graph Neural Networks^[57], and Diffusion models^[42] (see Supplementary Table S1). Notably, diffusion-based models (e.g., GenCast, FourCastNet3) enable native probabilistic sampling with explicit uncertainty representation, marking a shift from deterministic architectures where ensembles must be generated via input perturbations, often leading to spatial smoothing^[57].

WEATHER-CLIMATE HYBRIDS

Hybrid emulators are autoregressive ML models designed to bridge weather and climate timescales by incorporating external forcing, enabling long-term numerical stability while retaining the design of weather emulators. Although comparable to weather emulators in computational complexity, they aim to remain stable over much longer integrations. Examples of this emerging class include climaX^[54], ACE^[18,27,28], neuralGCM^[17], LUCIE^[62,66], CAMulator^[43], and cBottle^[48] (see Supplementary Table S2). Development efforts are increasingly targeting fully coupled ML emulators that jointly represent the atmosphere and ocean, exemplified by Ola^[70], DLESyM^[20], ACE2-SOM^[19],^[72].

Input and output Weather-climate hybrid emulators mirror the high spatio-temporal resolution and autoregressive design of weather models, but crucially incorporate time-varying external forcing, 12/13 approaches rely on SST, enabling Atmospheric Model Intercomparison Project (AMIP)-style experiments over extended timescales. Recent developments aim to expand their climate applicability by including additional anthropogenic forcing, such as CO₂ concentrations (e.g., ACE2^[18] or LUCIE-3D^[66]) or gridded emissions for historical projections (e.g., fine-tuning of climaX^[54]). Such developments ultimately move approaches closer to the climate world (Fig. 4).

Like their weather counterparts, hybrid emulators typically operate on high-dimensional three-dimensional grids with fine spatio-temporal resolution, e.g., ACE operates on $1^\circ \times 1^\circ$ horizontal resolution, and, apart from one model, all offer six-hourly temporal increments (see Supplementary Table S2). However, some approaches like LUCIE-2D and LUCIE-3D adopt a lighter-weight setup with coarser ($3.75^\circ \times 3.75^\circ$) grids^[62,66].

Training and evaluation Training datasets for hybrid emulators are highly mixed, utilizing either reanalysis data (e.g., ERA5 for neuralGCM^[17]) or direct outputs from physics-based models, ranging from idealised runs with annually repeating conditions (e.g., ACE^[27]) to historic simulations with realistic forcing (e.g., ACE2^[18]). Validation strategies occupy a middle ground, primarily focusing on long-term stability under annually repeating SST for hundreds to thousands of years^[18,27,28,74] or under realistic historic forcings over multi-decadal periods (10–80 years)^[17,27] (Fig. 3, panel b). Stability remains a primary bottleneck in hybrid emulation; for instance, NeuralGCM achieved stability in only 60% of 40-year simulations^[17], while the original

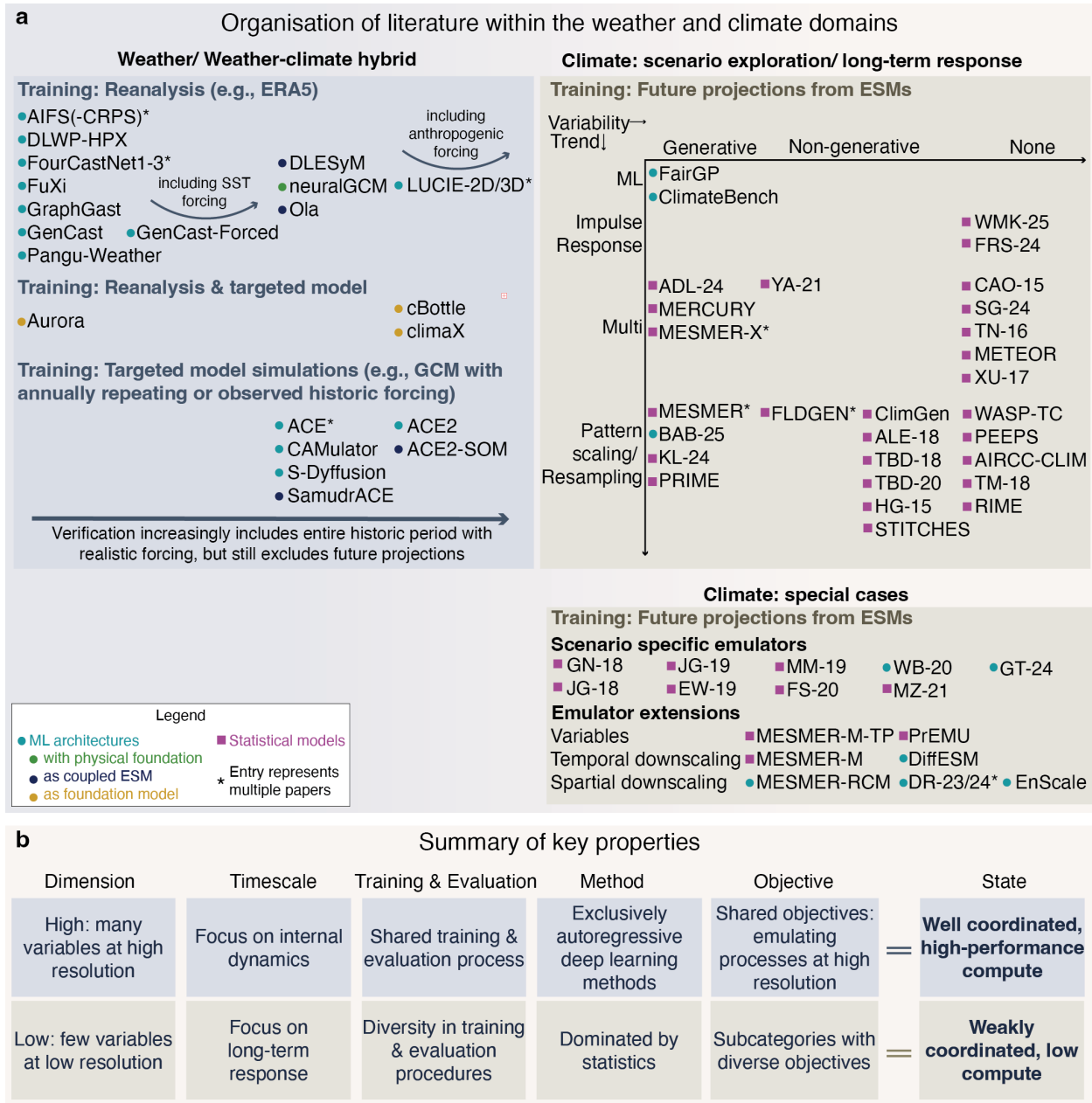


Figure 4. Data sources, design, and capabilities of weather, hybrid, and climate emulators. The emulator space is organized from left to right: Weather emulators (left), Hybrid emulators (middle), and Climate emulators (right). Data sources included during training (e.g., ERA5, CMIP future projections) are organized vertically. Emulator names (or paper shortcuts) are highlighted in color, with colors indicating whether an approach belongs to the machine-learning or statistical domain. The climate emulation space is broken down into substructures: approaches that take external forcing as input are organized according to how trend and variability are represented, and other approaches are categorized as either Scenario-Specific Emulators or Emulator Extensions. Overall, the weather and hybrid space is highly consolidated with well-defined and shared objectives, while the climate emulation space is methodologically and structurally heterogeneous.

ACE model exhibited a persistent cold bias^[27]. ACE2^[18] represents a significant advancement, achieving arbitrary stability over millennial timescales and partially mitigating this cold bias by incorporating CO₂ as

an explicit forcing. However, while stable, ACE2 still struggles to produce a fully realistic climate sensitivity to CO₂ increases. These generalization gaps are further highlighted by CAMulator, which encountered significant performance degradation when tasked with extrapolating to out-of-distribution regimes, such as +4K sea surface temperature (SST) shifts^[43]. Consequently, challenges in maintaining stability and physical consistency in the presence of historic forcing trends persist, and none of these hybrid models have yet been evaluated in the context of future climate projections.

Methods Hybrid emulators generally build on the deep-learning methods used for weather emulators, including Spherical Fourier Neural Operators^[27], Neural Networks^[17], and Transformers^[43]. A key distinction and technical challenge is the need to ensure long-term stability and physical consistency over climate timescales. This is typically achieved in one of two ways: by incorporating differentiable solvers for partial differential equations (e.g., using a physics-based dynamical core as in NeuralGCM^[17]) alongside the ML components, or by explicitly enforcing physics-based conservation laws during training^[18,43]. This autoregressive, physics-informed design facilitates the development of fully coupled ML-based ESMs (e.g., ACE2-SOM^[19] and DLESyM^[20]), which are on the frontier of this domain.

CLIMATE EMULATORS

Climate emulators (or emulators of ESM outputs^[13]) are typically non-autoregressive, statistical models designed primarily for long-term projection tasks, prioritizing computational stability and fidelity over climate timescales under anthropogenic forcing (Fig. 3, panel b). Exceptions exist where statistical autoregression is used to introduce temporal correlation into the residual variability e.g.,^[11], which is functionally distinct from the state-dependent autoregression used by deep learning models to advance the full variable field. In this sense, these models emulate statistical outputs rather than dynamical processes, making them fundamentally distinct from weather or hybrid emulators as they are built to perform scenario experiments (e.g., ScenarioMIP^[99]). Well-established examples relying on statistical methods include FLDGEN^[67,69], MESMER^[11,51,78], STITCHES^[85], or METEOR^[80]. Recently, ML emulators have started emerging, including frameworks developed within ClimateBench^[21], FairGP^[63], and novel diffusion-based approaches^[44]. Given that this review included more than 40 climate emulators, we mention these approaches illustratively in the main paper and provide a complete overview of reviewed approaches in Supplementary Section 1.3.

Input and output The climate emulator space is defined by high input heterogeneity—ranging from raw emissions to gridded climatic variables—and highly constrained, low-dimensional outputs, typically limited to less than three variables at annual resolution.

Regarding inputs, approaches fall into three broad categories (see Supplementary Table S3). The majority (32/49) fall into Category 1: Direct Forcing, meaning they work off external forcing, which itself widely varies, from raw gridded emissions (e.g., Watson-Parris et al.^[21]) to aggregate global climate indicators such as GMT e.g.,^[11,85]. Moving toward aggregate inputs delegates complexities like carbon-cycle or radiative-transfer uncertainties to an upstream SCM. The remaining approaches are either Category 2: Emulator Extensions (8/49), which receive gridded inputs to enhance temporal resolution e.g.,^[31] or generate additional variables e.g.,^[56], or Category 3: Scenario-Specific Tools (9/49), which do not rely on external forcing as the scenario dependence is absorbed into the model formulation e.g.,^[30]. The latter tools are designed to extend the ensembles sizes of a specific scenarios, for infilling or as information compression tools.

Crucially, the output of spatially resolved emulators is limited to a small set of variables on two-dimensional grids. More than half (28/49) approaches generate output for only a single variable. The outputs often include temperature (40/49) or water-cycle variables such as precipitation (21/49). All reviewed climate emulators operate on a two-dimensional grid (no vertical levels), usually with coarse 2.5°x2.5° or native ESM resolution although higher resolution can be obtained using downscaling extensions. Furthermore, many operate at annual or coarser resolution (24/49), and more than 1/4th of approaches (13/49) do not represent natural variability.

Training and evaluation In contrast to weather and hybrid emulators, climate emulators are uniquely trained and evaluated on modelled projections from CMIP protocols, such as historic and future runs from the ScenarioMIP^[99]. The 38 emulation approaches that do not fall within category 3 (scenario-specific tools), are designed for scenario exploration, i.e., to explore scenarios other than the ones used for calibration, which necessitates splitting data based on scenario rather than calendar year. Evaluation strategies span a wide

spectrum of rigor, ranging from in-sample testing (where the emulator is tested on data seen during training, e.g.,^[79]) to more demanding tests of cross-scenario generalisability (where the emulator is tested on entirely unseen forcing pathways) e.g.,^[49].

The most common test of generalizability is cross-scenario, in-range testing (17/38), meaning the emulator is tested on an unseen scenario that remains within the calibration range seen during training. However, the most conservative and rigorous tests are those that probe the limits of extrapolation: cross-scenario, out-of-range (e.g., testing on higher warming not seen during training) or cross-scenario, structural (e.g., training on abrupt and testing on smooth emissions changes). Given this diversity, performance metrics across emulators are not directly comparable and relate to different levels of rigour. Notably, only Womack et al.^[105] explicitly evaluate emulator performance under temporary overshoot scenarios, highlighting an important gap in the existing literature with respect to exploring plausible emission futures^[7].

Methods The design philosophy of climate emulators reflects a heritage in climate diagnostics focused on stability and reliability over long timescales, resulting in a dominance of statistical methods over deep learning architectures.

Emulators that rely on external forcing often assume the climate signal splits into a long-term trend and internal variability (analogous to the partition in Eq. 1). Modelling the mean response is historically dominated by pattern scaling^[13,106]. While robust, pattern scaling cannot account for responses operating on different timescales, a limitation addressed by more complex impulse response formulations^[71,91]. Other approaches favour a middle ground, incorporating different response timescales through diverse forcing^[49] or more complex dependencies^[93]. ML offers flexibility for fitting the trend, but extrapolation risks are large, and pattern scaling has been shown to outperform ML methods when training data is limited^[107]. Consequently, ML-based models for the mean response remain rare (only two cases^[21,63]).

For modelling variability, methods are also heavily statistical, and 12/31 scenario-exploration approaches do not model it at all. Those that do primarily rely on non-generative resampling (8/31) or semi-parametric/parametric statistical methods (9/31). 3/31 approaches employ true ML for variability estimation, including a diffusion model^[44] that captures spatial correlations but lacks temporal memory.

The remaining categories employ methods tailored to their specific use: Scenario-Specific Emulators (Category 3) use statistical time-series models with scenario-dependence absorbed into the coefficients, while Emulator Extensions (Category 2) represent the most frequent use of ML, with approaches like Diffusion Models being applied for high-fidelity spatial and temporal downscaling (4/7 cases).

SYNTHESIS

The weather and hybrid emulation space appears highly consolidated with well-defined objectives and consistent evaluation practices. This space has rapidly evolved over the past five years in a top-down fashion, where the overall goal defines the research agenda: to accurately emulate atmospheric dynamics autoregressively and ensure numerical stability over extended rollout periods. In contrast, the climate emulation landscape has evolved as a pragmatic response to data needs since the 1990s, with publication volumes steadily increasing over the years. Consequently, the space has emerged in a bottom-up fashion, with inhomogeneous training and evaluation standards, but with the consistent goal to provide robust projections of the climate system’s statistical response to specific anthropogenic forcing scenarios.

Overall, the climate emulator space currently seems less affected by the broader AI revolution that has profoundly affected the weather and hybrid worlds, resulting in low-dimensional output and a dominance of statistical methods such as pattern scaling, resampling approaches, and parametric sampling strategies. While the lines between weather and climate timescales have blurred, there remains a critical, untouched boundary between emulators that are stable over climate timescales (hybrids) and true climate emulators (projection-ready models): the former class has yet to be used for projection experiments in the presence of strong external forcing, which is the defining task of climate modelling (Fig. 3, panel b). To do so, also nested approaches coupling weather and climate emulators can be envisioned, similar to the coupling of SCMs and climate emulators^[108].

Table 2. A Roadmap for Climate Emulator Development. Mapping the landscape of emulated data across four primary applications and a cross-cutting deployment layer. The table evaluates current operational status against technical requirements, identifying critical research priorities for next-generation climate services.

Application	Attribution	Future Scenario Exploration	Impact & Risk Assessment	Coupled Decision Modelling
Description	Quantify effects of specific emissions/actors via counterfactual pathways.	Explore climate outcomes under novel, policy-relevant scenarios.	Translate climate variables into sectoral and regional impacts.	Dynamically couple climate with socio-economic models (e.g., IAMs).
Requirements	Historical interpolation; physically consistent and detectable response to small perturbations.	Robust scenario extrapolation (within and outside warming ranges seen during training); reliability under structural shifts (e.g., overshoot).	High-resolution multivariate outputs; joint consistency across variables; coupling to impact models or emulating impacts directly.	Low-latency; stable feedback loops; iterative compatibility.
Status	Global SCMs routinely used within attribution frameworks, but spatial emulators are underutilised although they can meet the requirements.	Strong for standard transient scenarios; weak for overshoot and structurally novel pathways.	Effective for limited variables; limited output space; weak coupling to downstream impact models; some emulators focusing on impact indicators exist.	Technically feasible (partially) but not yet integrated into operational IAM frameworks.
Action	Enable observation-driven emulation; benchmark against established attribution methods; develop standards for applied and legal use.	Expand training sets to overshoot and abrupt/extreme emission changes; stress-test out-of-distribution performance.	Expand output space; integrate with weather-model advances; develop end-to-end emulation.	Build prototypes of coupled systems; analyse feedback stability; develop uncertainty handling in iterative frameworks.
Cross-cutting		Deployment, Ecosystem & On-demand Climate Services		
Description	Provide a robust operational software ecosystem and on-demand climate services to enable independent application of emulators across applications and sectors.			
Requirements	Standardized APIs, high-quality documentation, modular architecture; low-latency, flexible outputs; access without requiring local HPC.			
Status	Fragmented tool space; most applications conducted by model developers; AI accessibility paradox persists (fast models offer limited outputs, complex models require heavy compute).			
Action	Develop hierarchical model ecosystems; foster independent application layers; expand compute access; standardize interfaces; prioritize decision-relevant outputs; integrate platforms for on-demand use across stakeholders.			

FROM LITERATURE TO ACTION: APPLICATIONS AND DEPLOYMENT

Emulators are tools complementary to ESMs, designed for contexts where scalable, probabilistic exploration outweighs the need for full mechanistic fidelity. By providing large ensembles across diverse forcings, structural uncertainty and realizations of internal variability, emulators enable high-volume analysis that is computationally infeasible with ESMs alone and enhance the agility of the climate modelling landscape. Building on these capabilities, we identify key opportunities and challenges that we organise across two dimensions: real-world applications (what emulators are and could be used for) and deployment layers (how they can be integrated into workflows).

In terms of applications, we focus on four key opportunities where climate emulators provide added value: attribution, i.e., quantifying the effects of specific emissions or actors through counterfactual pathways; exploration of future scenarios, i.e., assessing climate outcomes under policy-relevant trajectories; impacts and risk assessment, i.e., translating climatic variables into sectoral and regional consequences; and coupled decision modelling, i.e., dynamically linking climate systems with socio-economic models, for example within integrated assessment frameworks (Table 2).

In terms of deployment, we focus on turning emulators developed within research frameworks into operational software ecosystems and on-demand climate services (Table 2).

APPLICATION: ATTRIBUTION

In the context of attribution, emulators add value by enabling the rapid exploration of diverse counterfactual pathways that represent actor-specific emission trajectories. These datasets are increasingly used in legal contexts to assess actor-specific liability for climate-related damages (e.g., liability of countries or carbon majors)^[109,110]. Furthermore, they play an expanding role in policy debates on climate finance at both national and international levels—particularly in discussions of who should bear the costs of damages from extreme events^[3,111,112]. This also has growing implications for the private sector, where investment decisions are increasingly shaped by potential litigation risks.

For most attribution problems, the required data must accurately reproduce historical trends and provide physically consistent, detectable responses to small perturbations within observed warming ranges. This makes attribution a primary use case for emulated data. In particular, emulators could facilitate end-to-end attribution frameworks, where specific emissions sources are directly linked to localized impacts within a single computational framework spanning emissions, climate response, and impact modelling. Such an approach would provide a unified computational framework capable of serving the needs of diverse stakeholders. It should be noted, however, that some attribution frameworks focus on attributing future warming^[113], where data requirements align more closely with those of future scenario exploration.

Status & Challenges While attribution studies using global^[113,114] or spatially resolved^[5,115] emulators exist, standard approaches—particularly for Extreme Event Attribution^[116]—rely primarily on data from numerical climate models. Consequently, spatially resolved emulators appear underutilised in applied attribution contexts, as well as in subsequent policy and legal analyses with^[5] being the only attribution study based on spatially explicit emulators. A primary challenge lies in the absence of standardised emulator-based attribution protocols, similar to protocols for ESM data from^[116] and a lack of observation-driven or observation-constrained emulated ensembles as suggested in^[117].

Research Agenda To bridge this gap, the community should actively include observational datasets during emulator development and explore observation-driven emulation to reduce biases inherited from numerical climate models. Crucially, the community needs to develop emulator-based attribution protocols and standardised benchmarking protocols. In addition, emulators should be developed as modular climate services. By providing stakeholders—including local governments, NGOs, and legal entities—with tools to generate context-specific attribution data, the community can move toward a decentralized, end-to-end attribution pipeline that better serves diverse stakeholder needs and reduces reliance on resource-intensive, case-by-case scientific analysis for generating attribution data.

APPLICATION: FUTURE SCENARIO EXPLORATION

To explore climate outcomes under novel, policy-relevant scenarios, emulators must be able to reliably extrapolate to unseen scenarios within and beyond their training domain and remain robust under structural shifts, such as rapid changes in forcing or temperature overshoot pathways.

Status & Challenges At present, emulators face a notable extrapolation barrier when applied to new scenarios^[21,44]. Most are trained on a limited set of standard transient scenarios and are rarely validated on structurally distinct pathways (see Figure S1), making it difficult to assess their performance outside the training domain. As a result, the boundaries of emulator fidelity remain poorly understood. While emulators are not expected to capture highly nonlinear regime shifts such as tipping points, they can—under appropriate conditions—represent more complex dynamics, such as peak-and-decline scenarios^[71,105]. This depends on the availability of structurally similar training data and whether the explored warming ranges remain within, or close to, those represented in the training set.

Research Agenda To address these limitations, the community should develop rigorous stress-testing protocols for out-of-distribution scenarios and advance foundational research on training and evaluation strategies for emulators. In particular, it remains an open question whether standard transient climate simulations constitute the most suitable training data when generalisability is the primary objective.

APPLICATION: IMPACT & RISK ASSESSMENT

Impact and risk assessment requires translating climate variables into sectoral and regional consequences, such as damages to infrastructure, agricultural productivity, or human health. In this context, emulators can provide large ensembles of climate outcomes that serve as inputs to impact models, enabling probabilistic risk analysis and exploration of adaptation strategies^[6]. Beyond this, there is growing interest in directly emulating impacts, bypassing intermediate climate variables to generate decision-relevant metrics^[118].

To support coupling to impact models, emulators must generate multivariate, high-resolution outputs that are physically consistent across variables and scales. This includes not only commonly emulated variables such as temperature, but also precipitation, wind, soil moisture, and extremes, often at spatial and temporal resolutions suitable for impact modelling^[119]. Crucially, these variables must be jointly consistent, as many impact models depend on their co-variability rather than on individual fields in isolation.

Status & Challenges Current emulators remain limited in this regard. Most focus on a narrow set of variables and operate at relatively coarse spatial resolution, which constrains their applicability for many impact models. While downscaling methods exist e.g.,^[35,46], they are typically developed for outputs from ESMs and may not be robust when applied to emulator-generated data with different error structures^[120]. As a result, there is a persistent mismatch between what emulators deliver and what impact assessment frameworks require.

More fundamentally, the challenge lies in representing the joint distribution of multiple climate variables in a physically consistent manner. Even when individual variables are emulated with reasonable accuracy, inconsistencies in their co-variability can propagate through impact models and lead to unreliable risk estimates. At the same time, developments in weather emulation and hybrid modelling are beginning to produce high-resolution, multivariate outputs, but these approaches are not yet widely integrated into climate-scale applications.

There are approaches to emulate impact indicators as opposed to variables^[83,118,121]. However, approaches target a limited set of impacts and the verification of these results proves difficult on account of a lack of training and testing data^[118].

Research Agenda Addressing these challenges requires a shift toward multivariate and high-resolution emulation frameworks that explicitly account for dependencies across variables. This includes developing methods to ensure joint physical consistency, as well as systematically evaluating emulator performance in downstream impact models rather than in isolation. Closer integration with advances in weather and hybrid modelling could help bridge current capability gaps, particularly in generating high-frequency and spatially resolved data. In addition, generating impact datasets that allow to explore end-to-end approaches that directly emulate impacts is an important next step, in particular for applications that do not require intermediate climate variables.

APPLICATION: COUPLED DECISION MODELLING

Emulators also have the potential to act as lightweight, computationally efficient climate components within coupled decision-making frameworks, such as Integrated Assessment Models (IAMs)^[122]. In these settings, climate responses must be evaluated iteratively as part of a feedback loop, where decisions influence emissions, emissions influence climate, and climate outcomes in turn affect subsequent decisions.

This application places distinct requirements on emulators. In addition to representing climate responses accurately, they must operate with low latency, remain stable under repeated calls, and produce smooth and consistent responses to changes in input trajectories. These properties are essential to ensure that feedbacks between climate and decision systems do not introduce numerical instabilities or artefacts into the optimisation process.

Status & Challenges While simplified climate modules are already embedded in many IAM, spatially resolved or multivariate emulation is typically not included. As a result, IAM miss to depict scenario-to-climate linkages where, for example, changes in climatic conditions can lead to the degradation of arable land which should in turn impact land-use policy and pricing mechanisms .

This gap reflects both technical and structural challenges. From a technical perspective, the iterative coupling of emulators and decision-models has not been studied yet, particularly with respect to stability and the propagation of uncertainty. From a structural perspective, development of emulators and decision models has largely proceeded in parallel, with limited integration between communities. This has led to a lack of tools specifically designed for coupled applications.

Research Agenda Advancing this application requires the development of emulators tailored to iterative coupling, with a focus on stability, smoothness, and computational efficiency. Prototype systems that integrate emulators into IAMs or related frameworks should be developed to assess feasibility and identify practical bottlenecks. In parallel, methods for handling and propagating uncertainty in coupled systems need to be established, ensuring that feedbacks do not amplify errors or produce misleading results. Finally, closer collaboration between emulator developers and the IAM community will be essential to align model capabilities with the requirements of decision-making frameworks and to enable the systematic integration of climate dynamics into policy-relevant analyses.

CROSS-CUTTING: DEPLOYMENT, OPERATIONAL ECOSYSTEM AND ON-DEMAND CLIMATE SERVICES

Across all application domains, a common limitation emerges: while many of the required capabilities are increasingly within reach, their practical use remains constrained by the absence of an operational software ecosystem. Fully exploiting the potential of climate emulators requires not only advances in model development, but also the ability to deliver accessible, flexible, and user-oriented climate information^[123,124].

At present, emulator applications are rare and most are conducted within a closed development loop, where the same groups that build models also apply them. As a result, tools are often tailored to research workflows rather than to the needs of external users, and independent application remains limited. This restricts both the uptake of emulators in decision-making contexts and their validation under real-world conditions. In practice, only a small number of platforms provide user-facing interfaces that allow non-experts to explore climate outcomes e.g.,^[33,125].

Bridging this gap requires the development of a broader operational ecosystem around emulation. This includes user-facing tools and APIs that enable on-demand access to climate information, standardized interfaces and documentation to facilitate interoperability, and workflows that translate emulator outputs into decision-relevant quantities rather than raw model fields. While such systems are envisioned^[126], they are not yet operational. At the same time, issues of computational accessibility must be addressed to avoid reinforcing existing inequalities, as more advanced models increasingly rely on resource-intensive infrastructures.

Developing such an ecosystem would enable a shift from isolated, study-specific applications toward a more decentralized and scalable use of climate emulators across policy, finance, and adaptation contexts. Without these structural advances, improvements in model capability alone are unlikely to translate into broader real-world impact.

DISCUSSION

Our review formalises the climate-emulation task and reveals a central tension in how agility is currently understood and pursued. Although the field has made rapid progress in computational speed, these gains have not translated into broader flexibility or application-readiness. This shortfall stems from a prevailing design philosophy in which emulators are built to reproduce the full gridded output structure of ESMs. In replicating the model rather than the user's information needs, the field has tended to converge on a narrow

set of tasks (typically gridded temperature time series), with only a limited number of studies expanding the objective or demonstrating concrete use-cases. As a result, agility is often conflated with speed, even though in practice it also involves the ability to generate context-appropriate outputs, integrate into diverse applications, and remain adaptable to a wide range of users and institutions. Recognising and designing for this broader notion of agility is essential if emulators are to evolve from fast surrogates into genuinely flexible and accessible interfaces for climate information.

At the same time, the deep methodological silos separating weather, hybrid, and climate emulators reflect contrasting interpretations of what, exactly, is being emulated. Weather and hybrid approaches focus on process emulation, learning the underlying dynamical operator through complex, autoregressive deep-learning architectures. Climate emulators, by contrast, emphasise statistical output emulation, approximating conditional distributions using non-autoregressive, stability-oriented methods. This leaves a critical boundary between the fields: the leap from long-term numerical stability to reliable extrapolation under strong anthropogenic forcing. Even the most advanced hybrid models have yet to demonstrate robustness for future scenarios central to climate policy. Thus, while AI-based dynamics offer promise for high-fidelity climate-scale simulations, their reliability under the conditions that matter most remains untested, underscoring the need for rigorous, shared evaluation standards.

Layered on top of these conceptual and methodological divides is a structural challenge: the immense computational cost and institutional concentration of state-of-the-art deep-learning tools. This creates an AI Accessibility Paradox. The concentration of cutting-edge models within institutions with industrial-scale HPC infrastructure risks undermining one of the core aspirations of climate emulation—democratizing access to climate information. If the most capable models remain usable only by a small number of well-resourced actors, computational equity erodes, and agility becomes a privilege rather than a shared benefit. Addressing this requires cultivating a hierarchy of emulators of varying complexity, and ensuring that insights from the most sophisticated tools can be transferred, distilled, or approximated in forms accessible to the broader research community.

Taken together, these patterns outline the key barriers to the future utility of climate emulators. Realizing agility in a meaningful and holistic sense will require more than accelerating computation: it will demand output flexibility aligned with user needs, methodological reliability across the full forcing spectrum, and equitable access to tools and development pathways. Achieving this will depend on community-wide commitments to clear task definitions, transparent and application-relevant benchmarks, and cross-domain methodological integration. With these foundations in place, emulators can transition from computational shortcuts to indispensable, democratizing interfaces between complex climate dynamics and real-world decision-making.

METHODS

This paper is informed by a literature review on regional emulation approaches. On September 30th 2025, we conducted a Scopus search for studies containing any of the terms “surrogat*,” “emulat*,” or “stochastic generat*” in combination with either “earth system model” or “climate model.” The exact search string was:

```
(TITLE-ABS-KEY((emulat* OR surrogat* OR "stochastic generat*")
AND ("earth system model" OR "earth-system model" OR "climate model")))
AND PUBYEAR > 2010).
```

The search yielded 784 studies. Titles and abstracts were subsequently screened using the ASReview (Automated Systematic Review) tool, which ranks papers by estimated relevance using ML and iteratively updates this ranking based on user feedback^[127,128]. Screening was terminated after the algorithm identified 20 consecutive papers as irrelevant. This process resulted in 84 potentially relevant studies, all of which were reviewed in depth.

During full-text review, several studies were excluded as out of scope and we narrowed the time horizon to the past decade. It also became evident that the distinction between weather, hybrid weather–climate, and climate emulation has increasingly blurred. Because the original search strategy did not explicitly capture the weather or hybrid domains, we complemented it by screening references cited in the initially identified ML studies, as well as in recently published deep learning emulators (e.g., AURORA, FourCastNet, and ACE). After this additional step, we identified 76 relevant studies. From these, we extracted information on output variables and their dimensionality, the number of jointly emulated outputs, temporal and spatial

resolution, domain coverage, whether natural variability was modelled generatively, dependence on external forcing, approaches for modelling trends and variability, the ability to represent overshoots, whether variability depends on forcing, and the datasets used for training, testing, and validation.

AUTHOR CONTRIBUTIONS

Sa.S. and L.G. conceived the study, Sa.S. performed the analysis and generated the figures with contributions from L.G., C.W., and S.I.S. Sa.S. wrote the manuscript with contributions from L.G., C.W., C.F.S., and S.I.S.

ACKNOWLEDGMENTS

We are very grateful to Dr. Claudia Tebaldi for generously sharing her work and findings with us, for valuable discussions of our study, and for her thoughtful feedback on the manuscript. This work greatly benefited from her insights and support.

Sa.S. acknowledges support by the German Federal Environmental Foundation (DBU).

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

- [1] IPCC WG1. *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 1 edition, July 2023. ISBN 978-1-009-15789-6. doi: 10.1017/9781009157896.
- [2] IPCC WG2. *Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 1 edition, June 2023. ISBN 978-1-009-32584-4. doi: 10.1017/9781009325844.
- [3] Ilan Noy, Michael Wehner, Dáithí Stone, Suzanne Rosier, Dave Frame, Kamoru Abiodun Lawal, and Rebecca Newman. Event attribution is ready to inform loss and damage negotiations. *Nature Climate Change*, 13(12):1279–1281, 2023. doi: 10.1038/s41558-023-01865-4.
- [4] Christopher W. Callahan and Justin S. Mankin. Carbon majors and the scientific case for climate liability. *Nature*, 640(8060):893–901, April 2025. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-025-08751-3.
- [5] Sarah Schöngart, Zebedee Nicholls, Roman Hoffmann, Setu Pelz, and Carl-Friedrich Schleussner. High-income groups disproportionately contribute to climate extremes worldwide. *Nature Climate Change*, 15(6):627–633, June 2025. ISSN 1758-678X, 1758-6798. doi: 10.1038/s41558-025-02325-x.
- [6] Sarah Schoengart, Moritz Schwarz, Jonas Schwaab, Felix Pretis, Peter Pfeleiderer, and Carl-Friedrich Schleussner. Temporary warming, lasting losses: Economic damages from climate overshoot may be irreversible within this century, August 2025.
- [7] Carl-Friedrich Schleussner, Gaurav Ganti, Quentin Lejeune, Biqing Zhu, Peter Pfeleiderer, Ruben Prütz, Philippe Ciais, Thomas L Frölicher, Sabine Fuss, Thomas Gasser, et al. Overconfidence in climate overshoot. *Authorea Preprints*, 2023.
- [8] Chao Li, Francis W. Zwiers, Xuebin Zhang, Erich M. Fischer, Fujun Du, Jieyu Liu, Jianyu Wang, Yongxiao Liang, Tong Li, and Lina Yuan. Constraining the entire earth system projections for more reliable climate change adaptation planning. *Science Advances*, 11(9):eadr5346, February 2025. ISSN 2375-2548. doi: 10.1126/sciadv.adr5346.
- [9] Venkatramani Balaji, Eric Maisonnave, Niki Zadeh, Bryan N. Lawrence, Joachim Biercamp, Uwe Fladrich, Giovanni Aloisio, Rusty Benson, Arnaud Caubel, Jeffrey Durachta, Marie-Alice Foujols, Grenville Lister, Silvia Mocavero, Seth Underwood, and Garrett Wright. Cpmip: Measurements of real computational performance of earth system models in cmip6. *Geoscientific Model Development*, 10(1): 19–34, January 2017. ISSN 1991-9603. doi: 10.5194/gmd-10-19-2017.
- [10] Mario C. Acosta, Sergi Palomas, Stella V. Paronuzzi Ticco, Gladys Utrera, Joachim Biercamp, Pierre-Antoine Bretonniere, Reinhard Budich, Miguel Castrillo, Arnaud Caubel, Francisco Doblas-Reyes, Italo Epicoco, Uwe Fladrich, Sylvie Joussaume, Alok Kumar Gupta, Bryan Lawrence, Philippe Le Sager,

- Grenville Lister, Marie-Pierre Moine, Jean-Christophe Rioual, Sophie Valcke, Niki Zadeh, and Venkatramani Balaji. The computational and energy cost of simulation and storage for climate science: Lessons from cmip6. *Geoscientific Model Development*, 17(8):3081–3098, April 2024. ISSN 1991-9603. doi: 10.5194/gmd-17-3081-2024.
- [11] Lea Beusch, Lukas Gudmundsson, and Sonia I. Seneviratne. Emulating earth system model temperatures with mesmer: From global mean temperature trajectories to grid-point-level realizations on land. *Earth System Dynamics*, 11(1):139–159, February 2020. ISSN 2190-4987. doi: 10.5194/esd-11-139-2020.
 - [12] Ivan Sudakow, Michael Pokojovy, and Dmitry Lyakhov. Statistical mechanics in climate emulation: Challenges and perspectives. *Environmental Data Science*, 1:e16, 2022. ISSN 2634-4602. doi: 10.1017/eds.2022.15.
 - [13] C. Tebaldi, N.E. Selin, R. Ferrari, and G. Flierl. Emulators of climate model output. *Annual Review of Environment and Resources*, June 2025. ISSN 1543-5938, 1545-2050. doi: 10.1146/annurev-environ-012125-085838.
 - [14] IPCC. Summary for policymakers. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou, 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*, pages 1–31. Cambridge University Press, Cambridge, UK and New York, NY, USA, 2021. doi: 10.1017/9781009157896.001.
 - [15] IPCC. Summary for policymakers. In H. O. Pörtner, D. C. Roberts, M. Tignor, E. S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, and B. Rama, editors, *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK and New York, NY, USA, 2022. doi: 10.1017/9781009325844.001.
 - [16] Intergovernmental Panel on Climate Change. Scoping of the ipcc seventh assessment report (ar7): Background, cross-cutting issues and the ar7 synthesis report. Technical Report IPCC-LXII/INF.7, Intergovernmental Panel on Climate Change, January 2025. URL <https://apps.ipcc.ch/eventmanager/documents/88/290120250857-INF.%207%20-%20Scoping%20of%20the%20AR7.pdf>.
 - [17] Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P. Brenner, and Stephan Hoyer. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, August 2024. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-024-07744-y.
 - [18] Oliver Watt-Meyer, Brian Henn, Jeremy McGibbon, Spencer K. Clark, Anna Kwa, W. Andre Perkins, Elynn Wu, Lucas Harris, and Christopher S. Bretherton. Ace2: Accurately learning subseasonal to decadal atmospheric variability and forced responses. *npj Climate and Atmospheric Science*, 8(1):205, May 2025. ISSN 2397-3722. doi: 10.1038/s41612-025-01090-0.
 - [19] Spencer K. Clark, Oliver Watt-Meyer, Anna Kwa, Jeremy McGibbon, Brian Henn, W. Andre Perkins, Elynn Wu, Lucas M. Harris, and Christopher S. Bretherton. 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*, 2(4):e2024JH000575, December 2025. ISSN 2993-5210, 2993-5210. doi: 10.1029/2024JH000575.
 - [20] Nathaniel Cresswell-Clay, Bowen Liu, Dale Durran, Zihui Liu, Zachary I. Espinosa, Raul Moreno, and Matthias Karlbauer. A deep learning earth system model for efficient simulation of the observed climate, February 2025.
 - [21] D. Watson-Parris, Y. Rao, D. Olivié, Ø. Seland, P. Nowack, G. Camps-Valls, P. Stier, S. Bouabid, M. Dewey, E. Fons, J. Gonzalez, P. Harder, K. Jeggle, J. Lenhardt, P. Manshausen, M. Novitasari, L. Ricard, and C. Roesch. Climatebench v1.0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems*, 14(10):e2021MS002954, October 2022. ISSN 1942-2466, 1942-2466. doi: 10.1029/2021MS002954.
 - [22] Cornel Soci, Hans Hersbach, Adrian Simmons, Paul Poli, Bill Bell, Paul Berrisford, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Raluca Radu, Dinand Schepers, Sebastien Villaume, Leopold Haimberger, Jack Woollen, Carlo Buontempo, and Jean-Noël Thépaut. The era5 global reanalysis from 1940 to 2022. *Quarterly Journal of the Royal Meteorological Society*, 150(764):4014–4048, October 2024.

ISSN 0035-9009, 1477-870X. doi: 10.1002/qj.4803.

- [23] Ronald Gelaro, Will McCarty, Max J. Suárez, Ricardo Todling, Andrea Molod, Lawrence Takacs, Cynthia A. Randles, Anton Darmenov, Michael G. Bosilovich, Rolf Reichle, Krzysztof Wargan, Lawrence Coy, Richard Cullather, Clara Draper, Santha Akella, Virginie Buchard, Austin Conaty, Arlindo M. Da Silva, Wei Gu, Gi-Kong Kim, Randal Koster, Robert Lucchesi, Dagmar Merkova, Jon Eric Nielsen, Gary Partyka, Steven Pawson, William Putman, Michele Rienecker, Siegfried D. Schubert, Meta Sienkiewicz, and Bin Zhao. The modern-era retrospective analysis for research and applications, version 2 (merra-2). *Journal of Climate*, 30(14):5419–5454, July 2017. ISSN 0894-8755, 1520-0442. doi: 10.1175/JCLI-D-16-0758.1.
- [24] Simon Lang, Mihai Alexe, Matthew Chantry, Jesper Dramsch, Florian Pinault, Baudouin Raoult, Mariana C. A. Clare, Christian Lessig, Michael Maier-Gerber, Linus Magnusson, Zied Ben Bouallègue, Ana Prieto Nemesio, Peter D. Dueben, Andrew Brown, Florian Pappenberger, and Florence Rabier. Aifs – ecmwf’s data-driven forecasting system, August 2024.
- [25] Gabriel Moldovan, Ewan Pinnington, Ana Prieto Nemesio, Simon Lang, Zied Ben Bouallègue, Jesper Dramsch, Mihai Alexe, Mario Santa Cruz, Sara Hahner, Harrison Cook, Helen Theissen, Mariana Clare, Cathal O’Brien, Jan Polster, Linus Magnusson, Gert Mertes, Florian Pinault, Baudouin Raoult, Patricia de Rosnay, Richard Forbes, and Matthew Chantry. An update to ecmwf’s machine-learned weather forecast model aifs, September 2025.
- [26] Simon Lang, Mihai Alexe, Mariana C. A. Clare, Christopher Roberts, Rilwan Adewoyin, Zied Ben Bouallègue, Matthew Chantry, Jesper Dramsch, Peter D. Dueben, Sara Hahner, Pedro Maciel, Ana Prieto-Nemesio, Cathal O’Brien, Florian Pinault, Jan Polster, Baudouin Raoult, Steffen Tietsche, and Martin Leutbecher. Aifs-crps: Ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score, 2024.
- [27] Oliver Watt-Meyer, Gideon Dresdner, Jeremy McGibbon, Spencer K. Clark, Brian Henn, James Duncan, Noah D. Brenowitz, Karthik Kashinath, Michael S. Pritchard, Boris Bonev, Matthew E. Peters, and Christopher S. Bretherton. Ace: A fast, skillful learned global atmospheric model for climate prediction, December 2023.
- [28] James P. C. Duncan, Elynn Wu, Jean-Christophe Golaz, Peter M. Caldwell, Oliver Watt-Meyer, Spencer K. Clark, Jeremy McGibbon, Gideon Dresdner, Karthik Kashinath, Boris Bonev, Michael S. Pritchard, and Christopher S. Bretherton. Application of the ai2 climate emulator to e3smv2’s global atmosphere model, with a focus on precipitation fidelity. *Journal of Geophysical Research: Machine Learning and Computation*, 1(3):e2024JH000136, September 2024. ISSN 2993-5210, 2993-5210. doi: 10.1029/2024JH000136.
- [29] Sameh Abdulah, Allison H. Baker, George Bosilca, Qinglei Cao, Stefano Castruccio, Marc G. Genton, David E. Keyes, Zubair Khalid, Hatem Ltaief, Yan Song, Georgiy L. Stenchikov, and Ying Sun. Boosting earth system model outputs and saving petabytes in their storage using exascale climate emulators. In *SC24: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–12, Atlanta, GA, USA, November 2024. IEEE. ISBN 979-8-3503-5291-7. doi: 10.1109/SC41406.2024.00008.
- [30] Matthew Edwards, Stefano Castruccio, and Dorit Hammerling. A multivariate global spatiotemporal stochastic generator for climate ensembles. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3):464–483, September 2019. ISSN 1537-2693. doi: 10.1007/s13253-019-00352-8.
- [31] Seth Basetti, Brian Hutchinson, Claudia Tebaldi, and Ben Kravitz. Diffesm: Conditional emulation of temperature and precipitation in earth system models with 3d diffusion models. *Journal of Advances in Modeling Earth Systems*, 16(10):e2023MS004194, October 2024. ISSN 1942-2466, 1942-2466. doi: 10.1029/2023MS004194.
- [32] Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the earth system. *Nature*, 641(8065):1180–1187, May 2025. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-025-09005-y.
- [33] Francisco Estrada, Oscar Calderón-Bustamante, Wouter Botzen, Julián A. Velasco, and Richard S.J. Tol. Aircc-clim: A user-friendly tool for generating regional probabilistic climate change scenarios and risk measures. *Environmental Modelling & Software*, 157:105528, November 2022. ISSN 13648152. doi:

10.1016/j.envsoft.2022.105528.

- [34] Geir-Arne Fuglstad and Stefano Castruccio. Compression of climate simulations with a nonstationary global spatiotemporal spde model. *The Annals of Applied Statistics*, 14(2), June 2020. ISSN 1932-6157. doi: 10.1214/20-AOAS1340.
- [35] Antoine Doury, Samuel Somot, Sebastien Gadat, Aurélien Ribes, and Lola Corre. Regional climate model emulator based on deep learning: Concept and first evaluation of a novel hybrid downscaling approach. *Climate Dynamics*, 60(5-6):1751–1779, March 2023. ISSN 0930-7575, 1432-0894. doi: 10.1007/s00382-022-06343-9.
- [36] Matthias Karlbauer, Nathaniel Cresswell-Clay, Dale R. Durran, Raul A. Moreno, Thorsten Kurth, Boris Bonev, Noah Brenowitz, and Martin V. Butz. Advancing parsimonious deep learning weather prediction using the healpix mesh. *Journal of Advances in Modeling Earth Systems*, 16(8):e2023MS004021, August 2024. ISSN 1942-2466, 1942-2466. doi: 10.1029/2023MS004021.
- [37] Stacey E. Alexeeff, Doug Nychka, Stephan R. Sain, and Claudia Tebaldi. Emulating mean patterns and variability of temperature across and within scenarios in anthropogenic climate change experiments. *Climatic Change*, 146(3-4):319–333, February 2018. ISSN 0165-0009, 1573-1480. doi: 10.1007/s10584-016-1809-8.
- [38] Joseph Guinness and Dorit Hammerling. Compression and conditional emulation of climate model output. *Journal of the American Statistical Association*, 113(521):56–67, January 2018. ISSN 0162-1459, 1537-274X. doi: kst.
- [39] Antoine Doury, Samuel Somot, and Sebastien Gadat. On the suitability of a convolutional neural network based rcm-emulator for fine spatio-temporal precipitation. *Climate Dynamics*, 62(9):8587–8613, September 2024. ISSN 0930-7575, 1432-0894. doi: 10.1007/s00382-024-07350-8.
- [40] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hasanzadeh, Karthik Kashinath, and Animashree Anandkumar. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators, February 2022.
- [41] Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, pages 1–11, Davos Switzerland, June 2023. ACM. ISBN 979-8-4007-0190-0. doi: 10.1145/3592979.3593412.
- [42] Boris Bonev, Thorsten Kurth, Ankur Mahesh, Mauro Bisson, Jean Kossaifi, Karthik Kashinath, Anima Anandkumar, William D. Collins, Michael S. Pritchard, and Alexander Keller. Fourcastnet 3: A geometric approach to probabilistic machine-learning weather forecasting at scale, 2025.
- [43] William E. Chapman, John S. Schreck, Yingkai Sha, David John Gagne II, Dhamma Kimpara, Laure Zanna, Kirsten J. Mayer, and Judith Berner. Camulator: Fast emulation of the community atmosphere model, April 2025.
- [44] Shahine Bouabid, Andre Nogueira Souza, and Raffaele Ferrari. Score-based generative emulation of impact-relevant earth system model outputs, October 2025.
- [45] Andreas Groth and Erik Chavez. Efficient inference and learning of a generative model for enso predictions from large multi-model datasets. *Climate Dynamics*, 62(6):5259–5282, June 2024. ISSN 0930-7575, 1432-0894. doi: 10.1007/s00382-024-07162-w.
- [46] Maybritt Schillinger, Maxim Samarin, Xinwei Shen, Reto Knutti, and Nicolai Meinshausen. Enscale: Temporally-consistent multivariate generative downscaling via proper scoring rules, September 2025.
- [47] Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6(1):190, November 2023. ISSN 2397-3722. doi: 10.1038/s41612-023-00512-1.
- [48] Noah D. Brenowitz, Tao Ge, Akshay Subramaniam, Peter Manshausen, Aayush Gupta, David M. Hall, Morteza Mardani, Arash Vahdat, Karthik Kashinath, and Michael S. Pritchard. Climate in a bottle: Towards a generative foundation model for the kilometer-scale global atmosphere, July 2025.
- [49] Long Cao, Govindasamy Bala, Meidi Zheng, and Ken Caldeira. Fast and slow climate responses to CO_2 and solar forcing: A linear multivariate regression model characterizing transient climate change. *Journal of Geophysical Research: Atmospheres*, 120(23), December 2015. ISSN 2169-897X, 2169-8996. doi: 10.1002/2015JD023901.

- [50] Jaehong Jeong, Stefano Castruccio, Paola Crippa, and Marc G. Genton. Reducing storage of global wind ensembles with stochastic generators. *The Annals of Applied Statistics*, 12(1), March 2018. ISSN 1932-6157. doi: 10.1214/17-AOAS1105.
- [51] Shruti Nath, Quentin Lejeune, Lea Beusch, Sonia I. Seneviratne, and Carl-Friedrich Schleussner. Mesmer-m: An earth system model emulator for spatially resolved monthly temperature. *Earth System Dynamics*, 13(2):851–877, April 2022. ISSN 2190-4987. doi: 10.5194/esd-13-851-2022.
- [52] Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, January 2025. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-024-08252-9.
- [53] Bobby Antonio, Kristian Strommen, and Hannah M. Christensen. Seasonal forecasting using the gencast probabilistic machine learning model, 2025.
- [54] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. Climax: A foundation model for weather and climate, December 2023.
- [55] Jaehong Jeong, Yuan Yan, Stefano Castruccio, and Marc G. Genton. A stochastic generator of global monthly wind energy with tukey g-and-h autoregressive processes. *Statistica Sinica*, 2019. ISSN 10170405. doi: 10.5705/ss.202017.0474.
- [56] Sarah Schöngart, Lukas Gudmundsson, Mathias Hauser, Peter Pfleiderer, Quentin Lejeune, Shruti Nath, Sonia Isabelle Seneviratne, and Carl-Friedrich Schleussner. Introducing the mesmer-m-tpv0.1.0 module: Spatially explicit earth system model emulation for monthly precipitation and temperature. *Geoscientific Model Development*, 17(22):8283–8320, November 2024. ISSN 1991-9603. doi: 10.5194/gmd-17-8283-2024.
- [57] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, December 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adi2336.
- [58] Timothy J. Osborn, Craig J. Wallace, Ian C. Harris, and Thomas M. Melvin. Pattern scaling using climgen: Monthly-resolution future climate scenarios including changes in the variability of precipitation. *Climatic Change*, 134(3):353–369, February 2016. ISSN 0165-0009, 1573-1480. doi: 10.1007/s10584-015-1509-9.
- [59] M. Maimone, S. Malter, J. Rockwell, and V. Raj. Transforming global climate model precipitation output for use in urban stormwater applications. *Journal of Water Resources Planning and Management*, 145(6):04019021, June 2019. ISSN 0733-9496, 1943-5452. doi: 10.1061/(ASCE)WR.1943-5452.0001071.
- [60] Hao Pan, Lukas Gudmundsson, Mathias Hauser, Jonas Schwaab, Yann Quilcaille, and Sonia I. Seneviratne. Mesmer-rcm: A probabilistic climate emulator for regional warming projections, July 2025.
- [61] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, July 2023. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-023-06185-3.
- [62] Haiwen Guan, Troy Arcomano, Ashesh Chattopadhyay, and Romit Maulik. Lucie: A lightweight uncoupled climate emulator with long-term stability and physical consistency. *Journal of Advances in Modeling Earth Systems*, 17(11):e2025MS005152, November 2025. ISSN 1942-2466, 1942-2466. doi: 10.1029/2025MS005152.
- [63] Shahine Bouabid, Dino Sejdinovic, and Duncan Watson-Parris. Fairgp: A bayesian energy balance model for surface temperatures emulation. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS003926, June 2024. ISSN 1942-2466, 1942-2466. doi: 10.1029/2023MS003926.
- [64] Lőrinc Mészáros, Frank Van Der Meulen, Geurt Jongbloed, and Ghada El Serafy. A bayesian stochastic generator to complement existing climate change scenarios: Supporting uncertainty quantification in marine and coastal ecosystems. *Stochastic Environmental Research and Risk Assessment*, 35(3):719–736, March 2021. ISSN 1436-3240, 1436-3259. doi: 10.1007/s00477-020-01935-5.
- [65] Gang Liu, Shushi Peng, Chris Huntingford, and Yi Xi. A new precipitation emulator (premu v1.0) for lower-complexity models. *Geoscientific Model Development*, 16(4):1277–1296, February 2023. ISSN 1991-9603. doi: 10.5194/gmd-16-1277-2023.
- [66] Haiwen Guan, Troy Arcomano, Ashesh Chattopadhyay, and Romit Maulik. Lucie-3d: A three-

- dimensional climate emulator for forced responses, September 2025.
- [67] Robert Link, Abigail Snyder, Cary Lynch, Corinne Hartin, Ben Kravitz, and Ben Bond-Lamberty. Fldgen v1.0: An emulator with internal variability and space–time correlation for earth system models. *Geoscientific Model Development*, 12(4):1477–1489, April 2019. ISSN 1991-9603. doi: 10.5194/gmd-12-1477-2019.
 - [68] Theodore Weber, Austin Corotan, Brian Hutchinson, Ben Kravitz, and Robert Link. Technical note: Deep learning for creating surrogate models of precipitation in earth system models. *Atmospheric Chemistry and Physics*, 20(4):2303–2317, February 2020. ISSN 1680-7324. doi: 10.5194/acp-20-2303-2020.
 - [69] Abigail Snyder, Robert Link, Kalyn Dorheim, Ben Kravitz, Ben Bond-Lamberty, and Corinne Hartin. Joint emulation of earth system model temperature-precipitation realizations with internal variability and space-time and cross-variable correlation: Fldgen v2.0 software description. *PLOS ONE*, 14(10):e0223542, October 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0223542.
 - [70] Chenggong Wang, Michael S. Pritchard, Noah Brenowitz, Yair Cohen, Boris Bonev, Thorsten Kurth, Dale Durran, and Jaideep Pathak. Coupled ocean-atmosphere dynamics in a machine learning earth system model, June 2024.
 - [71] Lyssa M. Freese, Paolo Giani, Arlene M. Fiore, and Noelle E. Selin. Spatially resolved temperature response functions to co₂ emissions. *Geophysical Research Letters*, 51(15):e2024GL108788, August 2024. ISSN 0094-8276, 1944-8007. doi: 10.1029/2024GL108788.
 - [72] James P. C. Duncan, Elynn Wu, Surya Dheeshjith, Adam Subel, Troy Arcomano, Spencer K. Clark, Brian Henn, Anna Kwa, Jeremy McGibbon, W. Andre Perkins, William Gregory, Carlos Fernandez-Granda, Julius Busecke, Oliver Watt-Meyer, William J. Hurlin, Alistair Adcroft, Laure Zanna, and Christopher Bretherton. Samudrace: Fast and accurate coupled climate modeling with 3d ocean and atmosphere emulators, September 2025.
 - [73] Nadja Herger, Benjamin M. Sanderson, and Reto Knutti. Improved pattern scaling approaches for the use in climate impact studies. *Geophysical Research Letters*, 42(9):3486–3494, May 2015. ISSN 0094-8276, 1944-8007. doi: 10.1002/2015GL063569.
 - [74] Salva Rühling Cachay, Brian Henn, Oliver Watt-Meyer, Christopher S. Bretherton, and Rose Yu. Probabilistic emulation of a global climate model with spherical dyffusion, November 2024.
 - [75] Sarah Wilson Kemsley, Timothy J. Osborn, Stephen R. Dorling, and Craig Wallace. Pattern scaling the parameters of a markov-chain gamma-distribution daily precipitation generator. *International Journal of Climatology*, 44(1):144–159, January 2024. ISSN 0899-8418, 1097-0088. doi: 10.1002/joc.8320.
 - [76] Shruti Nath, Julie Carreau, Kai Kornhuber, Peter Pflöderer, Carl-Friedrich Schleussner, and Philippe Naveau. Mercury: A fast and versatile multi-resolution based global emulator of compound climate hazards, December 2024.
 - [77] Lea Beusch, Zebedee Nicholls, Lukas Gudmundsson, Mathias Hauser, Malte Meinshausen, and Sonia I. Seneviratne. 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*, 15(5):2085–2103, March 2022. ISSN 1991-9603. doi: 10.5194/gmd-15-2085-2022.
 - [78] Y. Quilcaille, L. Gudmundsson, L. Beusch, M. Hauser, and S. I. Seneviratne. Showcasing mesmer-x: Spatially resolved emulation of annual maximum temperatures of earth system models. *Geophysical Research Letters*, 49(17):e2022GL099012, September 2022. ISSN 0094-8276, 1944-8007. doi: 10.1029/2022GL099012.
 - [79] Yann Quilcaille, Lukas Gudmundsson, and Sonia I. Seneviratne. Extending mesmer-x: A spatially resolved earth system model emulator for fire weather and soil moisture. *Earth System Dynamics*, 14(6):1333–1362, December 2023. ISSN 2190-4987. doi: 10.5194/esd-14-1333-2023.
 - [80] Marit Sandstad, Norman Julius Steinert, Susanne Baur, and Benjamin Mark Sanderson. Meteorv1.0.1: A novel framework for emulating multi-timescale regional climate responses. *Geoscientific Model Development*, 18(21):8269–8312, November 2025. ISSN 1991-9603. doi: 10.5194/gmd-18-8269-2025.
 - [81] Ben Kravitz and Abigail Snyder. Pangeo-enabled esm pattern scaling (peeeps): A customizable dataset of emulated earth system model output. *PLOS Climate*, 2(12):e0000159, December 2023. ISSN 2767-3200. doi: 10.1371/journal.pclm.0000159.
 - [82] Camilla Mathison, Eleanor J. Burke, Gregory Munday, Chris D. Jones, Chris J. Smith, Norman J. Steinert, Andy J. Wiltshire, Chris Huntingford, Eszter Kovacs, Laila K. Gohar, Rebecca M. Varney, and Douglas McNeall. A rapid-application emissions-to-impacts tool for scenario assessment: Probabilistic regional impacts from model patterns and emissions (prime). *Geoscientific Model Development*, 18(5):

- 1785–1808, March 2025. ISSN 1991-9603. doi: 10.5194/gmd-18-1785-2025.
- [83] Edward Byers, Michaela Werning, Mahé Perrette, Niklas Schwind, Volker Krey, Keywan Riahi, and Carl-Friedrich Schleussner. Fast climate impact emulation for global temperature scenarios with the rapid impact model emulator (rime). *Environmental Research: Climate*, 4(3):035011, September 2025. ISSN 2752-5295. doi: 10.1088/2752-5295/adee3d.
 - [84] Yan Song, Zubair Khalid, and Marc G. Genton. Efficient stochastic generators with spherical harmonic transformation for high-resolution global climate simulations from cesm2-lens2. *Journal of the American Statistical Association*, 119(548):2493–2507, October 2024. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2024.2360666.
 - [85] Claudia Tebaldi, Abigail Snyder, and Kalyn Dorheim. Stitches: Creating new scenarios of climate model output by stitching together pieces of existing simulations. *Earth System Dynamics*, 13(4):1557–1609, November 2022. ISSN 2190-4987. doi: 10.5194/esd-13-1557-2022.
 - [86] Claudia Tebaldi and Reto Knutti. Evaluating the accuracy of climate change pattern emulation for low warming targets. *Environmental Research Letters*, 13(5):055006, May 2018. ISSN 1748-9326. doi: 10.1088/1748-9326/aabef2.
 - [87] C Tebaldi, A Armbruster, H P Engler, and R Link. Emulating climate extreme indices. *Environmental Research Letters*, 15(7):074006, July 2020. ISSN 1748-9326. doi: 10.1088/1748-9326/ab8332.
 - [88] Giang T. Tran, Kevin I. C. Oliver, András Sóbester, David J. J. Toal, Philip B. Holden, Robert Marsh, Peter Challenor, and Neil R. Edwards. Building a traceable climate model hierarchy with multi-level emulators. *Advances in Statistical Climatology, Meteorology and Oceanography*, 2(1):17–37, April 2016. ISSN 2364-3587. doi: 10.5194/ascmo-2-17-2016.
 - [89] Matthew A. Thomas and Ting Lin. A dual model for emulation of thermosteric and dynamic sea-level change. *Climatic Change*, 148(1-2):311–324, May 2018. ISSN 0165-0009, 1573-1480. doi: 10.1007/s10584-018-2198-y.
 - [90] Philip Goodwin, Martin Leduc, Antti-Ilari Partanen, H. Damon Matthews, and Alex Rogers. A computationally efficient method for probabilistic local warming projections constrained by history matching and pattern scaling, demonstrated by wasp-lgrtc-1.0. *Geoscientific Model Development*, 13(11):5389–5399, November 2020. ISSN 1991-9603. doi: 10.5194/gmd-13-5389-2020.
 - [91] Christopher B. Womack, Glenn Flierl, Shahine Bouabid, Andre N. Souza, Paolo Giani, Sebastian D. Eastham, and Noelle E. Selin. A framework for assessing and understanding sources of error in earth system model emulation, August 2025.
 - [92] Yangyang Xu and Lei Lin. Pattern scaling based projections for precipitation and potential evapotranspiration: Sensitivity to composition of ghgs and aerosols forcing. *Climatic Change*, 140(3-4):635–647, February 2017. ISSN 0165-0009, 1573-1480. doi: 10.1007/s10584-016-1879-7.
 - [93] Xiao-Chen Yuan, Nan Zhang, Wei-Zheng Wang, and Yi-Ming Wei. Large-scale emulation of spatio-temporal variation in temperature under climate change. *Environmental Research Letters*, 16(1):014041, January 2021. ISSN 1748-9326. doi: 10.1088/1748-9326/abd213.
 - [94] K. Hasselmann. Stochastic climate models part i. theory. *Tellus*, 28(6):473–485, December 1976. ISSN 00402826, 21533490. doi: 10.1111/j.2153-3490.1976.tb00696.x.
 - [95] Gregory M. Flato. Earth system models: An overview. *WIREs Climate Change*, 2(6):783–800, November 2011. ISSN 1757-7780, 1757-7799. doi: 10.1002/wcc.148.
 - [96] T. N. Palmer. A nonlinear dynamical perspective on climate prediction. *Journal of Climate*, 12(2):575–591, February 1999. ISSN 0894-8755, 1520-0442. doi: 10.1175/1520-0442(1999)012<0575:ANDPOC>2.0.CO;2.
 - [97] Michael J. McPhaden, Stephen E. Zebiak, and Michael H. Glantz. Enso as an integrating concept in earth science. *Science*, 314(5806):1740–1745, December 2006. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1132588.
 - [98] Shipra Jain, Adam A. Scaife, Theodore G. Shepherd, Clara Deser, Nick Dunstone, Gavin A. Schmidt, Kevin E. Trenberth, and Thea Turkington. Importance of internal variability for climate model assessment. *npj Climate and Atmospheric Science*, 6(1):68, June 2023. ISSN 2397-3722. doi: 10.1038/s41612-023-00389-0.
 - [99] Brian C. O’Neill, Claudia Tebaldi, Detlef P. Van Vuuren, Veronika Eyring, Pierre Friedlingstein, George Hurtt, Reto Knutti, Elmar Kriegler, Jean-Francois Lamarque, Jason Lowe, Gerald A. Meehl, Richard Moss, Keywan Riahi, and Benjamin M. Sanderson. The scenario model intercomparison project

- (scenariomip) for cmip6. *Geoscientific Model Development*, 9(9):3461–3482, September 2016. ISSN 1991-9603. doi: 10.5194/gmd-9-3461-2016.
- [100] Malte Meinshausen, Sarah CB Raper, and Tom ML Wigley. Emulating coupled atmosphere-ocean and carbon cycle models with a simpler model, magicc6–part 1: Model description and calibration. *Atmospheric Chemistry and Physics*, 11(4):1417–1456, 2011. doi: 10.5194/acp-11-1417-2011.
 - [101] Christopher J. Smith, Piers M. Forster, Myles Allen, Nicholas Leach, Richard J. Millar, Giovanni A. Passerello, and Leighton A. Regayre. Fair v1.3: A simple emissions-based impulse response and carbon cycle model. *Geoscientific Model Development*, 11(6):2273–2297, June 2018. ISSN 1991-9603. doi: 10.5194/gmd-11-2273-2018.
 - [102] Z. R. J. Nicholls, M. Meinshausen, J. Lewis, R. Gieseke, D. Dommenges, K. Dorheim, C.-S. Fan, J. S. Fuglestad, T. Gasser, U. Gollücke, P. Goodwin, C. Hartin, A. P. Hope, E. Kriegler, N. J. Leach, D. Marchegiani, L. A. McBride, Y. Quilcaille, J. Rogelj, R. J. Salawitch, B. H. Samset, M. Sandstad, A. N. Shiklomanov, R. B. Skeie, C. J. Smith, S. Smith, K. Tanaka, J. Tsutsui, and Z. Xie. Reduced complexity model intercomparison project phase 1: Introduction and evaluation of global-mean temperature response. *Geoscientific Model Development*, 13(11):5175–5190, 2020. doi: 10.5194/gmd-13-5175-2020.
 - [103] Jonathan A. Weyn, Dale R. Durran, and Rich Caruana. Can machines learn to predict weather? using deep learning to predict gridded 500-hpa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, 11(8):2680–2693, August 2019. ISSN 1942-2466, 1942-2466. doi: 10.1029/2019MS001705.
 - [104] Christopher D. Roberts, Retish Senan, Franco Molteni, Souhail Boussetta, Michael Mayer, and Sarah P. E. Keeley. Climate model configurations of the ecmwf integrated forecasting system (ecmwf-ifs cycle 43r1) for highresmip. *Geoscientific Model Development*, 11(9):3681–3712, September 2018. ISSN 1991-9603. doi: 10.5194/gmd-11-3681-2018.
 - [105] Christopher B. Womack, Paolo Giani, Sebastian D. Eastham, and Noelle E. Selin. Rapid emulation of spatially resolved temperature response to effective radiative forcing. *Journal of Advances in Modeling Earth Systems*, 17(1):e2024MS004523, 2025. ISSN 1942-2466. doi: 10.1029/2024MS004523.
 - [106] Benjamin D Santer, Tom ML Wigley, Michael E Schlesinger, and John FB Mitchell. Developing climate scenarios from equilibrium gcm results. 1990.
 - [107] Björn Lütjens, Raffaele Ferrari, Duncan Watson-Parris, and Noelle E. Selin. The impact of internal variability on benchmarking deep learning climate emulators. *Journal of Advances in Modeling Earth Systems*, 17(8):e2024MS004619, August 2025. ISSN 1942-2466, 1942-2466. doi: 10.1029/2024MS004619.
 - [108] Nicholas J. Leach, Christopher D. Roberts, Matthias Aengenheyster, Daniel Heathcote, Dann M. Mitchell, Vikki Thompson, Tim Palmer, Antje Weisheimer, and Myles R. Allen. Heatwave attribution based on reliable operational weather forecasts. *Nature Communications*, 15(1):4530, May 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-48280-7.
 - [109] Christopher W. Callahan and Justin S. Mankin. Carbon majors and the scientific case for climate liability. *Nature*, 640(8060):893–901, April 2025. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-025-08751-3.
 - [110] Joana Setzer and Catherine Higham. Global trends in climate change litigation: 2025 snapshot. Report, Grantham Research Institute on Climate Change and the Environment, London, UK, 2025. URL <https://www.lse.ac.uk/granthaminstitute/wp-content/uploads/2025/06/Global-Trends-in-Climate-Change-Litigation-2025-Snapshot.pdf>. Accessed: April 2, 2026.
 - [111] Andrew D King, Michael R Grose, Joyce Kimutai, Izidine Pinto, and Luke J Harrington. Event attribution is not ready for a major role in loss and damage. *Nature Climate Change*, 13(5):415–417, 2023. doi: 10.1038/s41558-023-01651-2.
 - [112] Sarah Schoengart. Income-based attribution of extrem events v0.0.2, February 2025.
 - [113] Alexander Nauels, Johannes Gütschow, Matthias Mengel, Malte Meinshausen, Peter U Clark, and Carl-Friedrich Schleussner. Attributing nauels2019attributinglong-term sea-level rise to paris agreement emission pledges. *Proceedings of the National Academy of Sciences*, 116(47):23487–23492, 2019.
 - [114] Yann Quilcaille, Lukas Gudmundsson, Dominik L. Schumacher, Thomas Gasser, Richard Heede, Corina Heri, Quentin Lejeune, Shruti Nath, Philippe Naveau, Wim Thiery, Carl-Friedrich Schleussner, and Sonia I. Seneviratne. Systematic attribution of heatwaves to the emissions of carbon majors. *Nature*, 645(8080):392–398, September 2025. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-025-09450-9.
 - [115] Lea Beusch, Alexander Nauels, Lukas Gudmundsson, Johannes Gütschow, Carl-Friedrich Schleussner, and Sonia I. Seneviratne. Responsibility of major emitters for country-level warming and extreme

- hot years. *Communications Earth & Environment*, 3(1):7, January 2022. ISSN 2662-4435. doi: 10.1038/s43247-021-00320-6.
- [116] Geert Jan Van Oldenborgh, Karin Van Der Wiel, Sarah Kew, Sjoukje Philip, Friederike Otto, Robert Vautard, Andrew King, Fraser Lott, Julie Arrighi, Roop Singh, and Maarten Van Aalst. Pathways and pitfalls in extreme event attribution. *Climatic Change*, 166(1-2):13, May 2021. ISSN 0165-0009, 1573-1480. doi: 10.1007/s10584-021-03071-7.
 - [117] Lea Beusch, Lukas Gudmundsson, and Sonia I. Seneviratne. Crossbreeding cmip6 earth system models with an emulator for regionally optimized land temperature projections. *Geophysical Research Letters*, 47(15):e2019GL086812, August 2020. ISSN 0094-8276, 1944-8007. doi: 10.1029/2019GL086812.
 - [118] Niklas Schwind, Mahé Perrette, Edward Byers, Annika Högner, Quentin Lejeune, Tessa Möller, Zebedee Nicholls, Peter Pfliegerer, Sarah Schöngart, Michaela Werning, and Carl-Friedrich Schleussner. Rime-x v1.0: Combining simple climate models, earth system models, and climate impact models into a unified statistical emulator for regional climate indicators, February 2026.
 - [119] Katja Frieler, Jan Volkholz, Stefan Lange, Jacob Schewe, Matthias Mengel, María Del Rocío Rivas López, Christian Otto, Christopher P. O. Reyer, Dirk Nikolaus Karger, Johanna T. Malle, Simon Treu, Christoph Menz, Julia L. Blanchard, Cheryl S. Harrison, Colleen M. Petrik, Tyler D. Eddy, Kelly Ortega-Cisneros, Camilla Novaglio, Yannick Rousseau, Reg A. Watson, Charles Stock, Xiao Liu, Ryan Heneghan, Derek Tittensor, Olivier Maury, Matthias Büchner, Thomas Vogt, Tingting Wang, Fubao Sun, Inga J. Sauer, Johannes Koch, Inne Vanderkelen, Jonas Jägermeyr, Christoph Müller, Sam Rabin, Jochen Klar, Iliusi D. Vega Del Valle, Gitta Lasslop, Sarah Chadburn, Eleanor Burke, Angela Gallego-Sala, Noah Smith, Jinfeng Chang, Stijn Hantson, Chantelle Burton, Anne Gädeke, Fang Li, Simon N. Gosling, Hannes Müller Schmied, Fred Hattermann, Jida Wang, Fangfang Yao, Thomas Hickler, Rafael Marcé, Don Pierson, Wim Thiery, Daniel Mercado-Bettín, Robert Ladwig, Ana Isabel Ayala-Zamora, Matthew Forrest, and Michel Bechtold. Scenario setup and forcing data for impact model evaluation and impact attribution within the third round of the inter-sectoral impact model intercomparison project (isimip3a). *Geoscientific Model Development*, 17(1):1–51, January 2024. ISSN 1991-9603. doi: 10.5194/gmd-17-1-2024.
 - [120] Alfonso Hernanz, Carlos Correa, Juan-Carlos Sánchez-Perrino, Ignacio Prieto-Rico, Esteban Rodríguez-Guisado, Marta Domínguez, and Ernesto Rodríguez-Camino. On the limitations of deep learning for statistical downscaling of climate change projections: The transferability and the extrapolation issues. *Atmospheric Science Letters*, 25(2):e1195, February 2024. ISSN 1530-261X, 1530-261X. doi: 10.1002/asl.1195.
 - [121] Rose Z. Abramoff, Philippe Ciais, Peng Zhu, Toshihiro Hasegawa, Hitomi Wakatsuki, and David Makowski. Adaptation strategies strongly reduce the future impacts of climate change on simulated crop yields. *Earth’s Future*, 11(4):e2022EF003190, April 2023. ISSN 2328-4277, 2328-4277. doi: 10.1029/2022EF003190.
 - [122] Claudia Tebaldi, Brian O’Neill, and Edward Byers. Climate impacts in scenarios: Time to close the loop? *Earth’s Future*, 14(3):e2025EF007622, March 2026. ISSN 2328-4277, 2328-4277. doi: 10.1029/2025EF007622.
 - [123] Jörn Hoffmann, Peter Bauer, Irina Sandu, Nils Wedi, Thomas Geenen, and Daniel Thiemert. Destination earth – a digital twin in support of climate services. *Climate Services*, 30:100394, April 2023. ISSN 24058807. doi: 10.1016/j.cliser.2023.100394.
 - [124] Victoria M. Bauer, Mathias Hauser, Yann Quilcaille, Sarah Schöngart, Lukas Gudmundsson, and Sonia I. Seneviratne. Mesmer v1.0.0: Consolidating the modular earth system model emulator into a sustainable research software package, December 2025.
 - [125] Climate Analytics and PROVIDE Consortium. Climate risk dashboard. <https://climate-risk-dashboard.iiasa.ac.at/>, 2024. Version 0.2.32, developed under the EU Horizon 2020 PROVIDE project.
 - [126] Francisco J. Doblas-Reyes, Jenni Kontkanen, Irina Sandu, Mario Acosta, Mohammed Hussam Al Turjman, Ivan Alsina-Ferrer, Miguel Andrés-Martínez, Leo Arriola, Marvin Axness, Marc Batlle Martín, Peter Bauer, Tobias Becker, Daniel Beltrán, Sebastian Beyer, Hendryk Bockelmann, Pierre-Antoine Bretonnière, Sebastien Cabaniols, Silvia Caprioli, Miguel Castrillo, Aparna Chandrasekar, Suvarchal Cheedela, Victor Correal, Emanuele Danovaro, Paolo Davini, Jussi Enkovaara, Claudia Frauen, Barbara Früh, Aina Gaya Àvila, Paolo Ghinassi, Rohit Ghosh, Supriyo Ghosh, Iker González, Katherine Grayson,

- Matthew Griffith, Ioan Hadade, Christopher Haine, Carl Hartick, Utz-Uwe Haus, Shane Hearne, Heikki Järvinen, Bernat Jiménez, Amal John, Marlin Juchem, Thomas Jung, Jessica Kegel, Matthias Kelbling, Kai Keller, Bruno Kinoshita, Theresa Kiszler, Daniel Klocke, Lukas Kluft, Nikolay Koldunov, Tobias Kölling, Joonas Kolstela, Luis Kornblueh, Sergey Kosukhin, Aleksander Lacima-Nadolnik, Jeisson Javier Leal Rojas, Jonni Lehtiranta, Tuomas Lunttila, Anna Luoma, Pekka Manninen, Alexey Medvedev, Sebastian Milinski, Ali Omar Abdelazim Mohammed, Sebastian Müller, Devaraju Naryanappa, Natalia Nazarova, Sami Niemelä, Bimochan Niraula, Henrik Nortamo, Aleksi Nummelin, Matteo Nurisso, Pablo Ortega, Stella Paronuzzi, Xabier Pedruzo Bagazgoitia, Charles Pelletier, Carlos Peña, Suraj Polade, Himansu Pradhan, Rommel Quintanilla, Tiago Quintino, Thomas Rackow, Jouni Räisänen, Maqsood Mubarak Rajput, René Redler, Balthasar Reuter, Nuno Rocha Monteiro, Francesc Roura-Adserias, Silva Ruppert, Susan Sayed, Reiner Schnur, Tanvi Sharma, Dmitry Sidorenko, Outi Sievi-Korte, Albert Soret, Christian Steger, Bjorn Stevens, Jan Streffing, Jaleena Sunny, Luigi Tenorio, Stephan Thober, Ulf Tigerstedt, Oriol Tinto, Juha Tonttila, Heikki Tuomenvirta, Lauri Tuppi, Ginka Van Thienen, Emanuele Vitali, Jost Von Hardenberg, Ingo Wagner, Nils Wedi, Jan Wehner, Sven Willner, Xavier Yepes-Arbós, Florian Ziemer, and Janos Zimmermann. The destination earth digital twin for climate change adaptation, August 2025.
- [127] Rens Van De Schoot, Jonathan De Bruin, Raoul Schram, Parisa Zahedi, Jan De Boer, Felix Weijdem, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, Albert Harkema, Joukje Willemsen, Yongchao Ma, Qixiang Fang, Sybren Hindriks, Lars Tummers, and Daniel L. Oberski. An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3(2):125–133, February 2021. ISSN 2522-5839. doi: 10.1038/s42256-020-00287-7.
- [128] ASReview LAB developers. Asreview lab - a tool for ai-assisted systematic reviews (v2.0a3). Zenodo, November 2024. URL <https://doi.org/10.5281/zenodo.14051216>.
- [129] Sarah Schöngart. Mesmer-m-tp version 0.1.0, April 2024.
- [130] Claudia Tebaldi and Julie M. Arblaster. Pattern scaling: Its strengths and limitations, and an update on the latest model simulations. *Climatic Change*, 122(3):459–471, February 2014. ISSN 1573-1480. doi: 10.1007/s10584-013-1032-9.
- [131] Shruti Nath, Lukas Gudmundsson, Jonas Schwaab, Gregory Duveiller, Steven J. De Hertog, Suqi Guo, Felix Havermann, Fei Luo, Iris Manola, Julia Pongratz, Sonia I. Seneviratne, Carl F. Schleussner, Wim Thiery, and Quentin Lejeune. Timber v0.1: A conceptual framework for emulating temperature responses to tree cover change. *Geoscientific Model Development*, 16(14):4283–4313, July 2023. ISSN 1991-9603. doi: 10.5194/gmd-16-4283-2023.
- [132] Stefano Castruccio and Michael L. Stein. Global space–time models for climate ensembles. *The Annals of Applied Statistics*, 7(3), September 2013. ISSN 1932-6157. doi: 10.1214/13-AOAS656.
- [133] Stefano Castruccio and Joseph Guinness. An evolutionary spectrum approach to incorporate large-scale geographical descriptors on global processes. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 66(2):329–344, February 2017. ISSN 0035-9254, 1467-9876. doi: 10.1111/rssc.12167.
- [134] Matthias Katzfuss and Florian Schäfer. Scalable bayesian transport maps for high-dimensional non-gaussian spatial fields. *Journal of the American Statistical Association*, 119(546):1409–1423, April 2024. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2023.2197158.
- [135] Wenju Cai, Benjamin Ng, Tao Geng, Lixin Wu, Agus Santoso, and Michael J. McPhaden. Butterfly effect and a self-modulating el niño response to global warming. *Nature*, 585(7823):68–73, September 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2641-x.

Supplementary for “A review of spatially explicit Earth System Model emulators”

1 OVERVIEW OF REVIEWED LITERATURE

In this section, we provide additional information on the reviewed literature within the weather, weather–climate hybrid and climate domain that complements information provided in the main manuscript.

1.1 WEATHER EMULATORS

The weather-emulator domain is well-consolidated, with a shared objective across studies: to emulate atmospheric dynamics autoregressively with high fidelity and long-term numerical stability. With the exception of recent extensions to GenCast (GenCast-forced^[53]), these models operate exclusively autoregressive—no external forcing is prescribed—and, aside from Aurora^[32], which is conceived as a foundation model, training and evaluation relies on reanalysis data (mostly ERA5^[22]).

This strong consensus in task definition and data use made it straightforward to organise the literature into a common structure (Table S1). In total, the review included 13 individual publications across 8 emulator families.

1.2 HYBRID WEATHER-CLIMATE EMULATORS

Hybrid weather–climate emulators share the key architectural features with weather emulators, but differ in one central respect: they rely on external forcing—most commonly SST—to maintain stability over extended timescales. The depict climatic aspect to varying degrees, with LUCIE-2D/3D^[62,66] reflecting a step towards the climate-emulation domain, as it can be forced solely with CO₂ concentrations, and follows a relatively lightweight design. We summarize the 14 studies that fall within 11 emulator families in Table S2.

Because external forcing is incorporated directly, hybrid models exhibit substantially more diverse training and evaluation strategies (Table S2). Many are trained on reanalysis data such as ERA5 (e.g., neuralGCM^[36]; LUCIE-2D/3D^[62,66]; DLESyM^[20]), whereas others are trained on outputs from physics-based models—either idealised annually repeating experiments (e.g., ACE^[27,28]) or historically forced simulations (e.g., CAMulator^[43]; ACE2^[18]). Foundation-model strategies are beginning to emerge in this domain as well: cBottle^[48] and climaX^[54] are pretrained on combinations of reanalysis and model-simulated data and subsequently fine-tuned for downstream tasks—mirroring the design philosophy underpinning Aurora.

Validation practices occupy a middle ground between weather and climate emulation. Some frameworks test climate stability under annually repeating SST for hundreds to thousands of years^[18,27,28,74], whereas others examine stability under realistic historical SST forcing across decadal to multi-decadal periods (10–80 years)^[17,27]. Stability remains a central challenge: neuralGCM achieved stability in only 22 of 37 runs over a 40-year period with realistic forcing^[17]; ACE exhibits a strong cold bias^[27], partially mitigated by introducing CO₂ forcing; CAMulator shows difficulties under SST perturbations up to +4 K^[43]. Training and evaluation over longer, climate-relevant periods—such as LUCIE-3D’s 40-year historical CO₂-forced experiment—represent important steps toward true climate-scale emulation.

In summary, hybrid weather–climate emulators face two overarching challenges: maintaining stability over climate timescales and accurately reproducing historical climate trends. Notably, none of the reviewed models has yet been evaluated under future projection scenarios.

Supplementary Table S1. Weather emulators. Columns: Key feature of hybrid emulators, rows: reviewed models. AR-x indicates the order of the autoregressive process. List of example variables is non-exhaustive. TISR = total incoming solar radiation.

Reference	Name	Architecture	Spatial resolution (°)	Temporal resolution (h)	Example variables (non-exhaustive)	Training data and setting
Pathak et al. ^[40] , Kurth et al. ^[41] , Bonev et al. ^[42]	FourCastNet	AR-1; Adaptive Fourier Neural Operator (AFNO); Vision Transformer (ViT) backbone utilizing mixing in the spectral domain. (v3 incorporates a generative Diffusion-based decoder and prescribed forcing).	0.25x0.25, at up to 5 levels	6	Temperature, precipitation, wind, specific humidity, geopotential, mean sea-level pressure	ERA5; training: 1979-2015, validation: 2016, 2017, testing: 2018-end
Bi et al. ^[61]	Pangu-Weather	AR-1; 3D Vision Transformer (3D-ViT); Employs Earth-specific hierarchical shifting windows and a specialized 3D-absolute-positional encoding.	0.25x0.25, at up to 13 levels	1, 3, 6, 24	Temperature, precipitation, wind, specific humidity, geopotential, mean sea-level pressure	ERA5; training: 1979-2017, validation: 2019, testing: 2018

Lam et al. ^[57]	GraphCast	AR-2; Graph Neural Network (GNN); Multi-mesh Encoder-Processor-Decoder architecture using a deep message-passing processor on an icosahedral grid.	0.25x0.25, at up to 37 levels	6	Temperature, precipitation, wind, specific humidity, geopotential, mean sea-level pressure	ERA5; training: 1979-2015, validation: 2015, 2016, testing: froze protocol and retrained 4 variants of models on 1979-2017/2018/2019/2020 and tested on 2017/2018/2019/2020-2021, respectively
Chen et al. ^[47]	FuXi	AR-2; U-Transformer; Cascaded architecture combining U-Net styled downsampling with Vision Transformer blocks.	0.25x0.25, at up to 13 levels	6	Temperature, precipitation, wind, specific humidity, geopotential, mean sea-level pressure	ERA5; training: 1979-2015, validation: 2016, 2017, testing: 2018
Karlbauer et al. ^[36]	DLWP-HPX	AR-2; Convolutional Neural Networks; Operates on a cubed-sphere HEALPix (HPX) grid with multi-step loss functions and TISR conditioning.	HPX64 (1°x1°), no real vertical layers	3	Temperature, geopotential (only 6 variables in total)	ERA5; training: 1979-2012, validation: 2013-2016, testing: 2017-2018

Lang et al. ^[24] , Moldovan et al. ^[25] , Lang et al. ^[26]	AIFS, AIFS- CRPS	AR-2; Transformer- based Graph Neural Net- works; includes stationary forcing; AIFS- CRPS relies on CRPS-based loss function for ensemble forecasting	0.25x0.25, at up to 13 ver- tical levels	6	Temperature, geopo- tential, humidity, wind, mean sea-level pressure, precipitation	pre-training: ERA5 (training: 1979- 2020, validation/test: n/a, but has been running four times daily since Octo- ber 2023); fine-tuning: operational IFS NWP
Price et al. ^[52] (Antonio et al. ^[53] for an extension with forcing)	GenCast (GenCast- forced/ -prescribed)	AR-2; Con- ditional Dif- fusion Model; Probabilistic generative archi- tecture utilizing a latent-space diffusion process to produce high- fidelity ensemble forecasts.	0.25x0.25, at up to 13 lev- els	12	Temperature, precipi- tation, wind, specific humidity, geopotential, mean sea-level pressure	ERA5; training: 1979-2017, validation: 2018, testing: froze protocol, retrained on 1979-2018 and tested on 2019
Bodnar et al. ^[32]	Aurora	AR-2; Foun- dation Trans- former; Large- scale 3D Perceiver-like architecture pre-trained on diverse datasets with flexible multi-resolution tokenization.	0.1x0.1, at up to 13 levels	6	Temperature, precipi- tation, wind, specific humidity, geopotential, mean sea-level pressure	e.g., ERA5, MERRA-2 and historic model simulations plus additional datasets during fine-tuning; validation during pre-training: 2020, testing: 2022 and 2023

Supplementary Table S2. Hybrid weather-climate emulators. Columns: Key feature of hybrid emulators, rows: reviewed models. AR-x indicates the order of the autoregressive process. List of example variables is non-exhaustive. Whenever information is omitted, we could not find it in the reference.

Reference	Name	Forcing	Backbone	Spatial resolution (°)	Temporal resolution (h)	Example variables (non-exhaustive)	Training and evaluation
Kochkov et al. ^[17]	neuralGCM	SST, sea-ice concentration	AR-1; Hybrid differentiable solver; Spectral transform dynamical core integrated with learned ML-based physics parametrisations.	up to 0.7x0.7, at up to 32 levels	6	Temperature, wind, surface pressure, specific humidity	ERA5; training: 1979-2017, validation: 2018, testing: retrained from 1979-2019 and tested on 2020; version of model testing extrapolation to warmer weathers (train: 1979-2000; test: 2001-2022)
Watt-Meyer et al. ^[27] , Duncan et al. ^[28]	ACE	SST	AR-1; Spherical Fourier Neural Operator (SFNO); Spectral convolutions with cyclic/polar reflection padding and physical conservation constraints.	1x1, at up to 7 resp. 8 levels	6	Temperature, wind, surface pressure, specific total water, precipitation	Modelled data (FV3GFS) with annually repeating climatological SST and fixed GHG and aerosol concentrations (1982-2012 average), 11 ensemble members over 10 years; training: 10 ensemble members, validation: remaining ensemble member, testing: forcing with realistic SST over 1990-2020; adjusted to emulate another model (EAMv2) with annually repeating climatologies of SST and sea-ice (2005-2014 average), 62-years of data (after spin-up); training: first 41 years, validation: next 10 years, testing: last 10 year

Cachay et al. ^[74]	Spherical Dyffusion (S-Dyffusion)	Skin temperature of open ocean (similar to forcing with SST)	AR-1; Probabilistic diffusion model; U-Net backbone utilizing score-based generative modeling optimized for spherical geometry.	1x1, at up to 8 vertical levels	6	Temperature, wind, atmospheric pressure at surface, specific total water	Modelled data (FV3GFS) with annually repeating climatological SST and fixed GHG and aerosol concentrations (1982-2012 average), 11 ensemble members over 10 years; training: 10 ensemble members, validation: remaining ensemble member (same set-up as ACE); tested against existing emulators (e.g., ACE)
Watt-Meyer et al. ^[18]	ACE2	SST, global mean CO ₂ concentration	AR-1; Optimized SFNO; Derived from ACE with enhanced spectral filtering for multi-decadal climate stability.	1x1, at up to 8 levels	6	Temperature, wind, surface pressure, specific total water, precipitation	ERA5 (training: 1940-1995,2011-2019,1021-2022; validation: 1996-2000, testing: 2001-2010,2020) and modelled data from SHIELD with historic forcings (training: 1940-1995,2011-2021, validation: 1996-2000, testing: 2001-2010);
Chapman et al. ^[43]	CAMulator	SST	AR-1; WXFormer (Hierarchical CrossFormer); Multi-scale embedding with long-short distance attention and hard-coded mass-energy conservation constraints.	0.9x1.25, at up to 32 levels	6	Temperature, wind, surface pressure, specific total water, precipitation	CAM6 with observed historic forcing, training: 1979-2010, validation: 200, testing: 2012-2014; testing included biased SST forcing (+2 K and +4 K)
Guan et al. ^[62]	LUCIE-2D	TISR ¹	AR-1; Spherical Neural Operator (SNO); Direct spectral-domain convolutions via Spherical Harmonic Transforms (SHT) to eliminate polar singularities.	3.75 x 3.75, at up to 8 levels (although only data from 3 was used)	6	Temperature, wind, humidity, surface pressure	ERA5; trained over 10 years and tested on 40 years, however, exact train-test splits not provided

Guan et al. ^[66]	LUCIE-3D	CO ₂ concentration, SST (optional)	AR-1; 3D-SNO; Extension of LUCIE-2D incorporating vertical cross-attention and zonal-mean spectral conditioning.	3.75 x 3.75, at up to 8 levels (although only data from 3 was used)	6	Temperature, wind, humidity, surface pressure, total precipitation	ERA5; trained on 30 years of ERA5 reanalysis, no other details provided; testing included biased SST forcing (+2 K and +4 K)
Nguyen et al. ^[54]	climaX	CO ₂ concentration, SST (task-dependent); fine-tuned version uses gridded GHG and aerosol emissions	AR-1/Direct; Foundation Vision Transformer (ViT); Cross-variable attention mechanism with randomized tokenization for multi-source data ingestion.	up to 1.4x1.4, at up to 7 levels	6	Temperature, wind, surface pressure, specific total water, precipitation	CMIP6 (pre-training), ERA5 (fine-tuning); e.g., for ERA5 training: 1979-2015, validation: 2016, testing: 2017, 2018
Brenowitz et al. ^[48]	cBottle	SST	Non-AR; Latent Diffusion Model; Generative foundation model using Vector Quantized-Variational Autoencoder and Diffusion Transformers for direct atmospheric state sampling.	5km (HEALPix grid, 12.5M pixels), up to 90 levels	N/A directly samples states, but trained on hourly data	Temperature, wind, surface pressure, specific total water, precipitation, cloud variables	ICON-Sapphire global 5km simulation output (training: 2020-2024, validation: 2024-2025), ERA5 reanalysis (training: 1980-2017, validation: 2018)

¹TISR provides information on diurnal and seasonal cycles but does not contain climate-change signals. LUCIE-2D is effectively a weather emulator; we list it as hybrid mainly because of the LUCIE-3D extension. Other models also incorporate TISR, but we omit listing it when other forcing is present.

Wang et al. ^[70]	Ola	SST (atmosphere component); winds, pressure, temperature (ocean component); no external forcings	AR-1; Dual-branch SFNO; Coupled air-sea architecture utilizing latent feature-exchange layers for interface coupling.	0.25x0.25,	24 (atmosphere), 48 (ocean/coupling)	Temperature, wind, mean sea-level pressure, SST	ERA5 (atmosphere), training: 1980-2016, validation: few samples from 2017, testing: 2017 onwards; UFS replay (ocean), training: 1994-2016, testing: 2017 onwards UFS replay dataset (primary training) and ERA5 (ground truth); training: 1994-2016, testing: 2017 onwards
Cresswell-Clay et al. ^[20]	DLESyM	SST (atmosphere component), radiation and wind variables (ocean component); only run in equilibrium settings (no anthropogenic signal)	AR-1; Coupled Neural Network; Separate atmosphere and ocean branches (see ^[36]) using Residual/Feed-Forward networks coupled via exchange fluxes.	1x1, at selected levels	6 (atmosphere); 48 (ocean/-coupling)	Temperature, precipitation, wind, specific humidity, geopotential, mean sea-level pressure	Atmosphere model as in ^[36] ; ERA5 and Satellite data
Clark et al. ^[19]	ACE2-SOM	Global mean CO ₂ concentration	AR-1; Physical-ML hybrid; ML-based atmosphere (ACE2) coupled to a 1D numerical Slab Ocean Model via surface flux exchange.	$\approx 1^\circ \times 1^\circ$, at up to 8 levels	6	Temperature, wind, surface pressure, specific total water, precipitation, SST	Modelled data (GFDL SHIELD-SOM); trained on 50 years each of $1\times$, $2\times$, and $4\times CO_2$ equilibrium climates; tested on unseen $3\times CO_2$ and non-equilibrium change scenarios.

Duncan et al. ^[72]	SamudrACE	global mean CO ₂ con- centration (atmosphere component forced with SST; ocean component forced with atmosphere fluxes, e.g., wind stress or surface heat)	AR-1; Deep ML- ML coupling; ACE2 cou- pled to Samu- dra oceanic SFNO via high-frequency state-vector synchronization.	$\approx 1^\circ \times 1^\circ$, at up to 8 atmo- spheric and 19 oceanic vertical levels	6 (atmo- sphere); 120 (ocean/ coupling)	Temperature, wind stress, precipitation, surface fluxes, poten- tial temperature, sea surface height, sea-ice concentration	Modelled data (GFDL CM4 coupled global climate model); components are first trained offline and then finetuned during coupling; training: 150 years of the CM4 pre-industrial control run (CM4-PiControl), vali- dation: 40-year segment from the same CM4-piControl run (years 161- 1200), testing: unseen transient cli- mate change scenarios, including 45- year 1pctCO ₂ run and a 200-year abrupt4xCO ₂ run.
----------------------------------	-----------	---	---	--	---	--	---

1.3 CLIMATE EMULATORS

Climate emulators are computational tools designed to approximate the output of fully coupled ESMs at lower cost. They differ in design objectives, input–output configurations, and methodological approaches. We group the reviewed emulators into three categories (Table S3): first, ScenarioMIP-style emulators which are designed to extrapolate to scenarios beyond the training ones (32 papers); second, emulator extensions, which enhance existing frameworks by adding variables or increasing spatio-temporal resolution (8 papers); and third, scenario-specific emulators, which generate data for predefined scenarios, often focusing on variability or data compression (9 approaches).

Input and Output Inputs to climate emulators are heterogeneous, ranging from global aggregates to fully gridded fields. Among ScenarioMIP-style emulators, ClimateBench^[21] employs the most comprehensive input representation, using spatially resolved GHG and aerosol emissions allowing to potentially capture localized effects such as air pollution. Other approaches rely on global emissions (4/32), concentrations (2/32), or radiative forcing (3/32), though the dominant strategy is to use global climate indicators (22/32), such as GMT. When concentrations, radiative forcing, or global indicators are used, the emulator is typically coupled to an SCM to generate these inputs. This reflects a progressive outsourcing of physical complexity, whereby processes such as carbon cycle feedbacks and radiative transfer are handled by SCMs rather than the emulator itself. Emulator extensions, which aim to increase output dimensionality or resolution, generally take gridded climate fields as inputs—for example, spatially resolved temperature^[31,51,65,129]. In contrast, scenario-specific emulators typically operate without explicit external forcing, instead encoding scenario dependence directly within model structure or parameters (e.g.,^[30,55]). Table S4 summarizes the eight input types identified in the literature.

Outputs are usually limited to a small set of climate variables, most commonly temperature (40/49) and precipitation (20/49). Joint generative modelling remains uncommon: only six approaches produce multiple variables simultaneously, each with distinct trade-offs. For example, STITCHES^[85] can generate an arbitrary number of variables but is non-generative and requires extensive training data, whereas MESMER-M-TP^[56] generates only two variables but operates with a single training trajectory. Temporal resolutions range from hourly to annual, with 24/49 emulators operating at annual or coarser scales; none explicitly resolve vertical structure. Natural variability is frequently omitted (13/49) or incorporated in a non-generative manner (12/46), with only a limited subset of models producing fully generative variability.

A detailed overview of input–output characteristics across all reviewed emulators is provided in Table S5.

Training and Evaluation Training and evaluation strategies depend on the emulation objective. ScenarioMIP-style emulators typically partition data across scenarios to assess generalisation to unseen forcings, whereas scenario-specific emulators split along ensemble members to evaluate the representation of within-scenario variability. Table S6 defines five common data-splitting strategies identified in the literature, ranging from in-sample validation (training and testing on the same data) to structurally distinct cross-scenario setups—for example, training on scenarios with abrupt emissions changes and testing on scenarios with transient forcing pathways (called *cross-scenario*, *structural*). Figure S1 summarises the distribution of emulators across these strategies. From the 37 studies that explicitly list their training procedure, most (17/37) employ cross-scenario, in-range evaluations (i.e., testing on unseen scenarios within warming ranges seen during training), followed by within-scenario splits (9/37) (i.e., testing on unseen ensemble members for a scenario seen during training).

Reported performance is strongly conditioned on the chosen validation setup. In-sample evaluations tend to yield optimistic estimates, whereas out-of-range and structurally distinct tests provide more conservative and informative assessments of generalisation. In addition, the number and diversity of scenarios and ensemble members used during training can materially influence emulator performance.

Validation targets vary with the intended application and may include grid-cell distributions, spatial covariance structures, extremes, or spectral characteristics (e.g., for energy system or agricultural impact assessments). Overall, there is no standardised framework for training, data partitioning, or validation of climate emulators.

Methods Climate emulators originate from early pattern-scaling techniques developed in the 1990s to approximate global-to-local climate responses^[106], and have since evolved into more sophisticated frameworks. Methodological approaches can be broadly decomposed into representations of the forced response (trend) and internal variability, as summarised in Table S7.

Classical statistical methods remain dominant (39/49), while machine learning (ML) approaches (10/49) are still comparatively limited in scope and capability. Current ML applications are primarily used as extensions to existing frameworks—for example, in downscaling tasks^[31,35,39,46]. In ScenarioMIP-style settings, ML-based approaches often struggle to capture variability^[21], or are applied as generative corrections to pattern-scaled inputs^[44].

As a result, fully generative ML-based climate emulators remain underdeveloped, representing a key area for methodological advancement.

Supplementary Table S3. Categories of climate emulators. The table lists three common types of climate emulators.

Categories	
ScenarioMIP-style emulators	Designed to extrapolate to new scenarios, normally outside of the scenarios used for training; e.g., ^[11,85] .
Emulator extensions	Designed to extend existing approaches to represent additional variables, or enhance spatio-temporal resolution of the output; e.g., ^[31,56,65]
Scenario-specific emulators	Designed to generate data for a pre-defined scenario, useful for data compression or infilling when training data for a specific scenario is available, sometimes these approaches approximate only internal variability after removing external forcing; e.g., ^[30,34,55] .

Supplementary Table S4. Inputs of climate emulators. The table lists eight common types of input used for emulator development.

Inputs	
Emissions-based	Prescribe emissions (e.g., carbon dioxide, methane, aerosols). Most directly linked to human activity and emulate the full response chain (e.g., ^[21,71]).
Concentration-based	Condition on concentrations (e.g., carbon dioxide). Skip carbon cycle dynamics (e.g., ^[93]).
Radiative-forcing-based	Prescribe effective radiative forcing, skipping chemistry and energy balance steps (e.g., ^[91]).
Global-climate-based	Use large-scale states such as GMT or ocean heat uptake. Common in pattern scaling and related methods, connecting global to regional properties (e.g., ^[11,77,85,130]).
Regional-climate-based	Use gridded inputs to generate additional gridded outputs (e.g., precipitation from temperature) or to downscale coarse variables to higher resolution (e.g., ^[31,51,65?]).
Mixed	Combine multiple drivers (e.g., carbon dioxide concentrations, solar irradiance, and GMT) to capture different response timescales (e.g., ^[49]).
Other	Capture drivers outside the usual emissions-centric framing (e.g., tree cover changes ^[131]).
None	Any emulator that does not require external forcing, such as all scenario-specific designs (e.g., ^[30,55])

Supplementary Table S5. Inputs and outputs of the reviewed climate emulators. Columns: Key features of each climate emulator. The “Design category” column follows the definitions in Table S3, and the “Forcing category” column follows Table S4. The “Natural variability” column indicates whether natural variability is represented and, if so, whether it is modelled in a generative or non-generative manner. In the “# Variables (jointly)” column, the first value denotes the number of variables with which the approach has been demonstrated, while the value in parentheses indicates the number of variables that can be generated jointly. The “Temporal resolution” column reports the highest temporal resolution stated in the respective paper (e.g., if both annual and monthly data are produced, only the monthly resolution is listed).

	Name	Design category	Forcing category	Natural variability, generative	# Variables (jointly)	Temporal resolution
Goodwin et al. ^[90]	WASP-LGRTC-1.0 (WASP-TC)	Scenario exploration	GC	no	1 (1)	annual
Kravitz and Snyder ^[81]	PEEPS	Scenario exploration	GC	no	3 (1)	monthly
Estrada et al. ^[33]	AIRCC-Clim	Scenario exploration	GC	no	2 (1)	monthly
Thomas and Lin ^[89]	TM-18	Scenario exploration	GC	no	1 (1)	multi-year average
Byers et al. ^[83]	RIME	Scenario exploration	GC	no	at least 10 (1)	annual
Xu and Lin ^[92]	XU-17	Scenario exploration	GC	no	3 (1)	annual
Cao et al. ^[49]	CAO-15	Scenario exploration	mixed	no	6 (1)	annual
Sandstad et al. ^[80]	METEOR	Scenario exploration	emission	no	2 (1)	annual
Womack et al. ^[91]	WMK-25	Scenario exploration	radiative	no	1 (1)	annual
Freese et al. ^[71]	FRS-24	Scenario exploration	emission	no	1 (1)	annual
Watson-Parris et al. ^[21]	ClimateBench	Scenario exploration	emission (gridded)	multiple approaches, some include generative variability	4 (1)	annual
Tebaldi and Knutti ^[86]	TBD-18	Scenario exploration	GC	multiple approaches, some include non-generative variability	2 (1)	annual

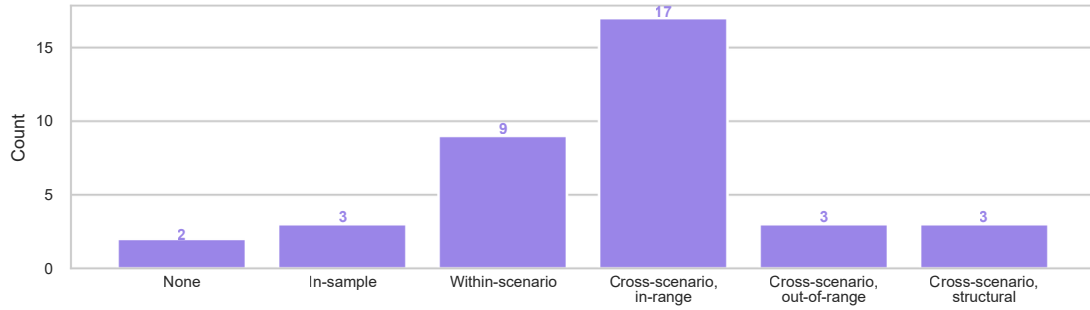
Tebaldi et al. ^[87]	TBD-20	Scenario exploration	GC	multiple approaches, include generative ability	ap-some non-vari-	at least 10 (1)	annual
Herger et al. ^[73]	HG-15	Scenario exploration	GC	multiple approaches, include generative ability	ap-some non-vari-	1 (1)	any
Mathison et al. ^[82]	PRIME	Scenario exploration	GC	non-generative		8 (1)	monthly
Osborn et al. ^[58]	ClimGen	Scenario exploration	GC	non-generative		1 (1)	monthly
Tebaldi et al. ^[85]	STITCHES	Scenario exploration	GC	non-generative		any number jointly, but only tested on 2 (2)	daily
Alexeeff et al. ^[37]	ALE-18	Scenario exploration	GC	non-generative		1 (1)	seasonal
Link et al. ^[67]	FLDGEN1.0	Scenario exploration	GC	pseudo-generative		1 (1)	annual
Snyder et al. ^[69]	FLDGEN2.0	Scenario exploration	GC	pseudo-generative		2 (2)	annual
Yuan et al. ^[93]	YA-21	Scenario exploration	concentration	pseudo-generative		1 (1)	annual
Song et al. ^[84]	SG-24	Scenario exploration	radiative	yes		1 (1)	daily
Beusch et al. ^[11]	MESMER	Scenario exploration	GC	yes		1 (1)	annual
Beusch et al. ^[77]	MESMER	Scenario exploration	GC	yes		1 (1)	annual
Quilcaille et al. ^[78]	MESMER-X	Scenario exploration	GC	yes		1 (1)	annual
Quilcaille et al. ^[79]	MESMER-X	Scenario exploration	GC	yes		4 (1)	annual
Bouabid et al. ^[44]	BAB-2025	Scenario exploration	GC	yes		4 (4)	monthly
Bouabid et al. ^[63]	FairGP	Scenario exploration	emission	yes		1 (1)	annual

Kemsley et al. ^[75]	KL-24	Scenario exploration	GC	yes	1 (1)	daily
Abdulah et al. ^[29]	ADL-24	Scenario exploration	radiative	yes	1 (1)	hourly
Tran et al. ^[88]	TN-16	Scenario exploration	concentration	no, but includes 90% of variability deterministically	1 (1)	annual
Nath et al. ^[76]	MERCURY	Scenario exploration	GC	yes	2 (2)	monthly
Jeong et al. ^[50]	JG-18	Scenario specific	none	yes	1 (1)	annual
Jeong et al. ^[55]	JG-19	Scenario specific	none	yes	1 (1)	monthly
Groth and Chavez ^[45]	GT-24	Scenario specific	none	models some of the variability	1 (1)	monthly
Edwards et al. ^[30]	EW-19	Scenario specific	none	yes	3 (3)	annual
Fuglstad and Castruccio ^[34]	FS-20	Scenario specific	none	yes	1 (1)	monthly
Guinness and Hammerling ^[38]	GN-18	Scenario specific	none	yes	1 (1)	single timestep
Weber et al. ^[68]	WB-20	Scenario specific	none	no	1 (1)	monthly
Mészáros et al. ^[64]	MZ-21	Scenario specific	none	no	1 (1)	daily
Maimone et al. ^[59]	MM-19	Scenario specific	none	yes	1 (1)	daily
Schöngart et al. ^[56]	MESMER-M-TP	Extension (variable)	RC	yes	2 (2)	monthly
Liu et al. ^[65]	PrEmu	Extension (variable)	RC	models some of the variability deterministically	2 (2)	monthly
Nath et al. ^[51]	MESMER-M	Extension (temporal resolution)	RC	yes	1 (1)	monthly
Bassetti et al. ^[31]	DiffESM	Extension (temporal resolution)	RC	yes	2 (1)	daily

Doury et al. ^[35]	DR-23	Extension (spatial resolution)	RC, concentration	non-generative	1 (1)	daily
Doury et al. ^[39]	DR-24	Extension (spatial resolution)	RC, concentration	non-generative	2 (1)	daily
Schillinger et al. ^[46]	EnScale	Extension (spatial resolution)	RC	yes	4 (4)	daily
Pan et al. ^[60]	MESMER-RCM	Extension (spatial resolution)	RC	yes	1 (1)	annual

Supplementary Table S6. Common strategies for testing emulator scenario-generalisability.

Strategy	Description and Examples
In-sample	Testing on samples the emulator was trained on. Example: Quilcaille et al. ^[78] train on all available ensemble members across scenarios and test on the same dataset.
Within-scenario	Testing on held-out samples from the same scenario. Examples: Song et al. ^[84] train on 20 ensemble members from SSP3-7.0 and test on the remaining 80; ? ^[1] train on a single ensemble member across scenarios and test on all other ensemble members under RCP8.5.
Cross-scenario, in-range	Testing on an unseen scenario with a GMT range within the forcing range seen during training. Example: Yuan et al. ^[93] train on RCP2.6 and RCP8.5, test on RCP4.5 and RCP6.0.
Cross-scenario, out-of-range	Testing on unseen scenarios outside the forcing range included during training. Example: Bouabid et al. ^[63] train on all scenarios except RCP8.5, then test on RCP8.5.
Cross-scenario, structural	Testing under scenarios that are structurally different from those used during training. Example: Castruccio and Stein ^[132] train on a scenario with an abrupt CO ₂ drop and test on one with gradually increasing concentrations.



Supplementary Figure S1. Data splitting for evaluation. Classification of emulators based on the splitting strategy employed for evaluation (see Table S6). We only include emulators explicitly designed for scenario exploration and report the most conservative test found in a study e.g., if an emulator performed *cross-scenario, out-of-range* tests as well as *cross-scenario, in-range*, we count the former test.

Supplementary Table S7. Overview of emulator methods. The top section summarises approaches for representing forced climate responses (trend), while the bottom section summarises approaches for representing variability.

Approaches for representing forced climate responses (trend)	
None	No explicit dependence on external forcing, i.e. $\hat{\mu}_t \equiv f(\mathbf{Y}_{<t}, \phi^{\text{TR}})$. Trends may be removed directly ^[50] or absorbed into autoregressive terms ^[30] .
Time-shift	Resampling approach that couples trend and variability. Data is reorganised by a predictive variable (e.g., GMT), generating new realisations by stitching together ESM samples aligned with a given forcing trajectory ^[73,85] .
Pattern scaling	GMT treated as predictor; f is a linear relationship estimated independently for each variable and location. Exploits near-linear scaling of local temperature with global warming ^[67,130] (see Fig. ??, Panel b).
Multi (response / time-scales)	Linear response models with additional predictors (e.g., ocean variables) or multiple time scales (e.g., distributed lag models), distinguishing fast and slow components of the response ^[73,93] .
Impulse response	Generalisation of linear response, representing responses to forcing as a convolution of impulse responses across time frames ^[91] .
ML	f inferred directly from data (e.g., neural networks, Gaussian processes, random forests). Some architectures (e.g., diffusion models) represent trend and variability jointly ^[21,63] .
Other	Approaches that do not fit other categories, e.g., with regional predictors ^[51,65?] .
Approaches for representing variability	
None	No variability is modelled; the emulator yields deterministic outputs ^[49,71] .
Resampling	Variability is generated by resampling. If the method decomposes into trend and variability, only residuals are resampled ^[37] ; otherwise, the full ESM output (trend + variability) can be resampled in one step ^[85] .
Parametric	Assumes a parameterised form for the residual distribution, e.g., Gaussian with a parameterised covariance matrix (such as distance-dependent covariance across grid cells) ^[84,133] .
Semi-parametric	Combines empirical information with parametric structure, e.g., smoothed empirical covariance matrices ^[11] .
Nonparametric	Fully data-driven approaches. For instance, Katzfuss and Schäfer ^[134] rely on transport maps; other possibilities include kernel density estimation.
ML	Generative ML models (e.g., diffusion models, probabilistic neural networks, normalising flows) used to sample variability. These typically model trend and variability jointly, without explicit decomposition ^[21,31] .

2 COMPUTATIONAL COMPLEXITY OF MODERN EMULATORS

Pangu-Weather consists of four deep networks with lead times of 1 h, 3 h, 6h and 24 h; Each of the four deep networks was trained for 100 epochs, and each of them takes approximately 16 days on a cluster of 192 NVIDIA Tesla-V100 GPUs. Each 3D deep network contains approximately 64 million parameters. Inference of a lead time of 56 h carries out a 24 h forecast, followed by a 6 h , 3h and 1h forecast so Pangu-Weather is designed to leverage each of the four networks;

3 NOTATION

Climatic variables such as temperature or precipitation are continuous fields defined over space and time. That is, a variable Y can be measured at any point in the three-dimensional spatial domain $D_s \subset \mathbb{R}^3$ (e.g., latitude, longitude, and height) and at any moment within a temporal domain $D_t \subset \mathbb{R}$. In practice, Y is measured or modelled at a discrete set of locations and times. For simplicity, we focus on surface-level variables at a fixed vertical height, so that each location $s \in D_s$ is indexed by a two-dimensional tuple of latitude, L , and longitude, l : $s = (L, l)$. We assume a regular spatial grid with k_L latitudes and k_l longitudes, yielding $N = k_L \cdot k_l$ total grid points, denoted $\mathbf{s} = \{(L_i, l_j)\}_{i=1, \dots, k_L}^{j=1, \dots, k_l}$. Time is indexed over n discrete steps $\mathbf{t} = \{t_i\}_{i=1, \dots, n}$.

Any measurement of a single climatic variable is a matrix $y \in \mathcal{Y}^{N \times n}$, where $\mathcal{Y} \subseteq \mathbb{R}$ reflects the valid range of the variable (e.g., $\mathcal{Y} = \mathbb{R}$ for temperature, $\mathbb{R}_{\geq 0}$ for precipitation, or $[0, 100]$ for relative humidity). The value of y at location s and time t is denoted $y_{s,t}$:

$$y = \begin{pmatrix} y_{s_1, t_1} & y_{s_2, t_1} & \cdots & y_{s_N, t_1} \\ y_{s_1, t_2} & y_{s_2, t_2} & \cdots & y_{s_N, t_2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{s_1, t_n} & y_{s_2, t_n} & \cdots & y_{s_N, t_n} \end{pmatrix}. \quad (5)$$

A column y_s gives the time series at location s , while a row y_t gives the spatial field at time t . To represent multiple variables—e.g., temperature, precipitation and wind—we define $\mathbf{y} \in \mathcal{Y}^{k \times N \times n}$, where k is the number of variables. The full dataset $\mathbf{y} = (y^{(1)}, \dots, y^{(k)})$ consists of k matrices, each describing one climatic variable over space and time. We denote $\mathbf{y}_t$ as the climatic state at time t .

We can best think about the climate state probabilistically. $\mathbf{y}_t$ is a measurement outcome that involved some randomness, i.e., if a butterfly had flapped its wings differently, we could have observed a different outcome, $\tilde{\mathbf{y}}_t$ [135]. This is referred to as deterministic chaos and we can mathematically capture it as apparent randomness, i.e., we promote the climate state to a random variable $\mathbf{Y}_t$ and interpret the actually measured climatic state as a sample

$$\mathbf{y}_t \sim \mathbf{Y}_t. \quad (6)$$

4 EMULATION OBJECTIVE

Ideally, the emulator parameters ϕ^{EMU} would be chosen to maximize the full joint likelihood of the climatic data evolving over time. Let $\mathbf{Y}_{1:T}^{\text{ESM}}$ reference the process evolving over time, i.e., $\mathbf{Y}_{1:T}^{\text{ESM}} \equiv \{Y_{t_0}, Y_{t_1}, \dots, Y_T\}$. Then we define the log-likelihood in terms of realised samples:

$$\hat{\phi}^{EMU} = \arg \max_{\phi^{EMU}} \log P(\mathbf{y}_{1:T}^{\text{ESM}} | F_{\leq T}, \phi^{EMU}) = \arg \max_{\phi^{EMU}} \sum_{t=1}^T \log P_t(\mathbf{y}_t^{\text{ESM}} | \mathbf{y}_{<t}^{\text{ESM}}, F_{\leq t}, \phi^{EMU}). \quad (7)$$

However, this is computationally not feasible and relies on assumptions that simplify the objective. A common simplification for building emulators is pattern scaling, which assumes that the spatiotemporal response of climate variables can be expressed as a linear function of a global or regional scalar (typically GMT). Formally, let ΔT_t denote the scalar temperature change at time t , and let $\mathbf{y}_t^{\text{ESM}}$ be the output of an ESM. Pattern scaling assumes that

$$\mathbf{y}_t^{\text{ESM}} \approx \mathbf{a} \Delta T_t + \mathbf{b} + \boldsymbol{\epsilon}_t, \quad (8)$$

where $\mathbf{a}$ and $\mathbf{b}$ are spatial patterns and offsets, respectively, and $\boldsymbol{\epsilon}_t$ is usually Gaussian noise.

Under this approximation, the emulator no longer seeks to reproduce the full conditional distribution $P(\mathbf{Y}_{1:T}^{\text{ESM}} | \mathbf{Y}_{<T}, F_{\leq T})$, but rather a simplified conditional distribution

$$\mathbf{Y}_t^{\text{EMU}} \sim P(\mathbf{Y}_t | \Delta T_t, \phi_{\text{PS}}^{\text{EMU}}), \quad (9)$$

where $\phi_{\text{PS}}^{\text{EMU}}$ parametrises the linear response patterns and residual variability. Consequently, the corresponding log-likelihood reduces to a sum over individual time steps, conditioned on ΔT_t :

$$\mathcal{L}_{\text{PS}}(\phi_{\text{PS}}^{\text{EMU}}) = \sum_{t=1}^T \log P(\mathbf{y}_t^{\text{ESM}} | \Delta T_t, \phi_{\text{PS}}^{\text{EMU}}). \quad (10)$$

This reduced formulation makes estimation computationally feasible and is widely used in climate impact studies. However, it implicitly assumes that the full spatiotemporal dependence of the climate fields is captured by the scalar ΔT_t , which may neglect nonlinear interactions, time-lagged dependencies, or internal variability not explained by the linear.

5 COMPLEXITY

Complexity here refers to the number of potential spatio-temporal cross-dependencies between variables, time steps, and locations in the covariance structure. We assume the emulator preserves cross-correlations among all variables at all time steps and all locations, and can be captured as a multivariate Gaussian process. We further assume the covariance matrix is symmetric (i.e., only the upper triangle, including the diagonal, is unique). An upper triangle matrix with k variables, over n timesteps and at N locations has $1/2 * (knN) * (knN + 1)$ free parameters, meaning its complexity is dominated by the leading order term $\mathcal{O}((knN)^2)$.