A Research Report on Artificial Intelligence in Climate Modeling

Author: Harshit Kumawat (Created using Agentic AI Pipeline)
Disclamer: This report is auto-generated by an AI-powered research assistant. Human verification is recommended for critical use.
Date of Creation: 2025-09-11

Abstract

This knowledge base synthesizes recent advances at the intersection of artificial intelligence and climate modeling, focusing on datasets, generative and conditional ML methods, natural language tools for climate information extraction, and operational/ethical considerations. ClimateSet provides a modular pipeline and a core ML-ready dataset aggregating outputs from 36 CMIP6 Earth System Models and associated Input4MIPs emissions to enable multi-model emulation and uncertainty-aware ML training. Generative approaches—particularly conditional diffusion models (DDPMs) with U-Net denoisers—have proven effective for high-factor precipitation downscaling, predicting noise via L2 objectives and producing multiple plausible high-resolution realizations for uncertainty quantification. NLP contributions include large questionnaire-derived datasets (CLIMA-CDP and CLIMA-INS) and CLIMABENCH, showing that self-supervised alignment methods can map unstructured reports to structured policy questionnaires and that general-purpose language models frequently outperform domain-specific variants on benchmarked tasks. Broader considerations include the environmental footprint of training/deploying large models, data provenance and IP issues, and the need to pair ML surrogates with multi-model ensembles to capture epistemic uncertainty. Collectively, these works indicate a promising role for AI to accelerate and scale climate science—provided careful attention to dataset harmonization, model evaluation, uncertainty communication, and resource impacts.

Methodology

This research report was generated using an **Agentic AI pipeline** designed to simulate the process of academic research, writing, and review. The methodology combines automated information retrieval, structured extraction, natural language generation, and iterative critique to ensure reliability and coherence. The pipeline consists of the following components:

1. **Searcher Agent**
   - Retrieves relevant Wikipedia articles, arXiv research papers, and recent news using specialized tools.
   - Ensures coverage of both academic and practical sources within a defined time period.
2. **Extractor Agent**
   - Processes the raw sources and converts them into a structured **knowledge base (JSON format)**.
   - Summarizes each topic and subtopic into concise bullet points with references.
3. **Writer Agent**
   - Expands the structured knowledge into detailed, human-readable sections.
   - Produces coherent paragraphs while maintaining alignment with the knowledge base.
4. **Critic Agent**
   - Reviews the Writer's output against the knowledge base.
   - Detects hallucinations, unsupported claims, or factual drift.
   - Provides corrective feedback or validates correctness.
5. **Assembler Agent**
   - Integrates all validated sections into a unified document.
   - Produces the final **PDF report** with a Title page, abstract, table of contents, Main body, conclusion, references, appendix, and consistent styling.

This layered methodology ensures that the generated report is **factually grounded, logically structured, and stylistically coherent**, while also being transparent about its AI-assisted origin.

Large-scale climate datasets and pipelines for ML (ClimateSet)

ClimateSet provides a consistent, machine-learning-ready, multi–climate-model dataset together with a modular pipeline that retrieves and preprocesses climate model inputs (Input4MIPs) and outputs (CMIP6) to support large-scale ML tasks in climate science [4]. The resource aggregates outputs and input emission fields to create harmonized input/output tensors suitable for model training and benchmarking, emphasizing monthly variables initially such as surface temperature and precipitation [4]. By packaging both a downloader and a preprocessor, ClimateSet aims to make reproducible dataset construction from CMIP6 and Input4MIPs straightforward for ML researchers while reducing the need for bespoke climate-domain preprocessing for each model [4].

The core composition of ClimateSet includes outputs from 36 Earth System Models (ESMs) together with associated emission fields across multiple Shared Socioeconomic Pathway (SSP) scenarios and historical runs, with an initial focus on key variables like temperature and precipitation [4]. These core data are publicly available and the release emphasizes ML readiness by harmonizing variable names and units, aligning spatial and temporal resolutions across models, and providing automated regridding and temporal aggregation to produce consistent tensors [4]. These design choices are intended to address common cross-model challenges such as inconsistent grid definitions, differing temporal resolutions, and unit heterogeneity so that models trained on the dataset can directly consume inputs without per-model adaptation [4].

A primary goal of ClimateSet is to supply sufficient, diverse training data to enable development of large ML climate models or “super emulators” that generalize across models and scenarios, rather than being tied to a single ESM [4]. Training across multiple models also allows ML practitioners to capture multi-model projection uncertainty, which is critical for policymaking and risk assessment and cannot be obtained from single-model datasets alone [4]. ClimateSet demonstrates its utility as a benchmark for climate model emulation on temperature and precipitation tasks, showing that multimodel-trained ML approaches reveal performance and generalization characteristics that single-model datasets cannot provide [4].

ClimateSet also acknowledges limitations in the core release: remaining inconsistencies across models persist (for example in ensemble member availability and vertical level definitions), the initial variable and scenario selection is limited, and the pipeline depends on the availability of CMIP6/Input4MIPs data and the ESGF infrastructure for retrieval [4]. The project mitigates some of these constraints by providing modular downloader and preprocessor components so users can extend the dataset to include additional variables, height levels, ensemble members, or scenarios as needed, while recognizing that some heterogeneities are inherent in the source model outputs [4].

Dataset composition and scope

The core release of ClimateSet aggregates outputs from 36 climate models and their associated emission inputs, focusing initially on monthly fields such as surface temperature and precipitation across several SSP scenarios as well as historical runs [4]. This aggregation is intended to create a multimodel corpus that supports ML tasks requiring extensive and diverse training data while preserving the link between forcings (Input4MIPs) and ESM responses (CMIP6) [4].

To enable direct ML training without bespoke preprocessing for each model, the dataset targets ML readiness by providing aligned grids, consistent units, and standardized temporal resolution across all included models [4]. These harmonization steps—unit conversion, temporal aggregation to standardized monthly cadence, and automated spatial regridding—produce uniform tensors that permit straightforward batching and model input pipelines typical in ML workflows [4].

Pipeline, tools and extensibility

ClimateSet includes a downloader that automates retrieval of CMIP6 and Input4MIPs files from the Earth System Grid Federation (ESGF) and a modular preprocessor that performs regridding, unit harmonization, and temporal aggregation to create harmonized ML-ready tensors [4]. Both components are modular by design so that researchers can add variables, additional vertical levels, ensemble members, or extra scenario runs without modifying the core pipeline architecture [4].

The pipeline is intended to lower domain-knowledge barriers for ML researchers by encapsulating climate-specific preprocessing steps and enabling reproducible dataset construction from the same underlying CMIP6/Input4MIPs sources [4]. By standardizing retrieval and preprocessing, the tools facilitate benchmarking and comparison of ML methods across consistent multimodel datasets while allowing extensibility to expand the dataset's scope over time [4].

Use cases and implications for climate ML

Primary ML use cases for ClimateSet include climate model emulation (developing fast surrogates for long-running simulations), statistical downscaling, scenario projection, extreme-event prediction, and training large multimodel emulators or "super emulators" that can quickly generate projections for new or perturbed scenarios [4]. The dataset's multimodel construction enables these applications to leverage greater data diversity and sample size, aligning with ML best practices for training large-capacity models [4].

Training across multiple climate models also facilitates explicit quantification of model-projection uncertainty, which is a critical input for policy decisions and risk assessment and cannot be captured by single-model emulators alone [4]. By providing a standardized benchmark for emulation on variables such as temperature and precipitation, ClimateSet allows researchers to evaluate both predictive skill and the robustness of ML models' generalization across ESMs and scenarios, informing the development of methods that are useful for decision-relevant climate applications [4].

Generative ML methods for climate downscaling and emulation

Generative machine learning models — including generative adversarial networks (GANs), variational autoencoders (VAEs), diffusion models, and conditional transformers — have been applied to climate tasks such as statistical downscaling (super-resolution), emulation (surrogate modeling of Earth system model components), and the generation of stochastic weather or climate realizations for uncertainty-aware applications [5,4,1]. Within this spectrum, diffusion-based generative models, specifically denoising diffusion probabilistic models (DDPMs) conditioned on low-resolution climate variables, have shown notable promise for high upscaling factors (e.g., $4\times$, $8\times$) in regional precipitation downscaling, providing superior fidelity relative to regression baselines and competitive or improved behavior versus GANs in terms of training stability and preservation of high-frequency spatial details [5].

Key architectural and algorithmic choices that support effective downscaling with these models include U-Net style backbones for the denoising networks, concatenation of upsampled low-resolution conditional inputs with high-resolution stochastic states, and loss formulations that directly predict the injected noise (epsilon) using an L2 objective as in the DDPM framework [5]. The stochastic nature of conditional generative approaches allows sampling of multiple plausible high-resolution realizations from the same low-resolution conditioning, enabling quantification of aleatoric uncertainty at local scales — a capability not available in deterministic regression-based downscalers [5]. In parallel, emulation benefits from large, multi-model datasets (for example, ClimateSet), which enable machine-learned surrogates to learn cross-model response patterns and generalize across forcing scenarios, thereby drastically reducing computational cost relative to full Earth system model runs while retaining important aspects of model variability [4,1].

Conditional diffusion model for precipitation downscaling

The conditional downscaling methodology adapts the DDPM framework to map low-resolution climate inputs X_LR to high-resolution precipitation fields X_HR by formulating a reverse diffusion chain p_theta(X_HR^{t-1} | X_HR^t, X_LR). Each reverse step is modeled as a Gaussian with a learned mean mu_theta and variance sigma_t^2, enabling iterative refinement from a noisy initial state toward a conditioned high-resolution sample [5]. During training, the forward (noising) process q is defined by a schedule of noise variances beta_t that progressively corrupt the high-resolution field; the neural denoiser epsilon_theta — commonly instantiated with a U-Net backbone — is trained to predict the injected Gaussian noise at each timestep using an L2 objective L = E[||epsilon - epsilon_theta(X_LR, X_HR^t, t)||^2] [5].

Implementation details that have proven effective include conditioning on multiple low-resolution variables (e.g., TS, PRECT, dPHIS), upsampling those low-resolution inputs to the high-resolution grid for concatenation with the stochastic state at each denoising step, and performing T iterative reverse steps to generate high-resolution realizations. This conditional DDPM approach has been demonstrated on CESM simulation data over North America, showing its ability to produce plausible high-resolution precipitation fields conditioned on coarse inputs [5].

Comparative methods, evaluation and robustness

Comparative evaluations indicate that diffusion-based downscaling models tend to outperform regression-based super-resolution approaches, particularly at large upscaling factors where regression methods commonly produce oversmoothed outputs, and they often exhibit more stable training dynamics than GAN-based alternatives [5,4,1]. Robust model assessment therefore requires a multi-axis evaluation protocol that combines pixel-wise skill metrics (e.g., RMSE, MAE), distributional diagnostics (such as precipitation frequency and intensity distributions), spatial-structure measures, and out-of-sample generalization tests across different time slices or forcing scenarios to assess temporal and scenario robustness [5,4,1].

Practical deployment considerations include the selection of conditioning variables, approaches for maintaining temporal consistency across sequential samples, and the computational cost associated with iterative sampling in diffusion methods. These trade-offs must be balanced against the benefits of uncertainty quantification: ensembles of stochastic samples permit estimation of aleatoric uncertainty at local scales, while training across multiple climate models or forcing scenarios helps capture epistemic uncertainty and improves generalization [5,4,1].

Natural language methods for climate information extraction and benchmarking

Large volumes of unstructured climate reports — including corporate, city, state, and national disclosures — create a pressing need for natural language processing (NLP) tools that can convert free-form text into structured questionnaires and fields useful for policy analysis, monitoring, and research [6]. These unstructured narratives contain policy-relevant information that is difficult to aggregate at scale without automated methods for aligning passages to predefined question schemas and extracting standardized responses [6].

To address this need, the CLIMA work introduces two large questionnaire-derived datasets and a broader benchmark suite that together enable systematic development and evaluation of climate-specific NLP methods [6]. The approach leverages existing semi-structured questionnaires to produce training signals and assembles CLIMABENCH, a compilation of climate text classification tasks, to provide standardized evaluation across topic classification, question–answer alignment, and report-to-question mapping tasks [6].

The proposed methods and resources are intended for practical applications such as automating questionnaire population, accelerating evidence aggregation for policy decisions, detecting greenwashing, and producing datasets suitable for fine-tuning large language models for climate policy tasks [6]. Empirically, the work also highlights that model choice and pretraining strategy materially affect performance: contrary to expectation, well-tuned general-purpose language models sometimes outperform domain-specific models (for example, ClimateBERT) on the compiled benchmark tasks, underscoring the importance of comprehensive evaluation when adapting models to climate document tasks [6].

Datasets and benchmark (CLIMA-CDP, CLIMA-INS, CLIMABENCH)

CLIMA-CDP is derived from Carbon Disclosure Project (CDP) disclosure questionnaires and provides broad coverage spanning thousands of organizations and many questionnaire items, enabling large-scale alignment between corporate disclosures and structured questions [6]. CLIMA-INS is constructed from insurance disclosures, specifically NAIC insurance climate risk survey responses, and after preprocessing yields roughly 17,000 question–answer pairs that can be used to train and evaluate passage-to-question alignment methods [6].

CLIMABENCH aggregates multiple climate NLP datasets, including the questionnaire-derived collections, to provide a standardized evaluation framework across a range of tasks relevant to climate information extraction. The benchmark supports comparisons on topic classification, question–answer alignment, and mapping unstructured reports to questionnaire fields, facilitating reproducible assessment of model capabilities on climate-specific challenges [6].

Modeling strategy and empirical findings

The modeling strategy centers on self-supervised classification that exploits the structure of existing questionnaires to create supervision signals: text passages are aligned to questionnaire questions, enabling in-domain training without the need for costly manual annotation [6]. Models trained in this manner were evaluated on a variety of tasks, including in-domain classification, cross-domain transfer, and real-world mapping tasks that align unstructured reports to questionnaire fields; experiments demonstrate that this approach supports generalization across stakeholder types such as cities, corporations, and states [6].

Empirical results from the CLIMA evaluations reveal an important nuance in domain adaptation: domain-tailored pretraining is not always superior to using well-tuned general-purpose models. In several compiled benchmark tasks, general models outperformed specialized domain models (for example, ClimateBERT), indicating that pretraining strategy and downstream tuning are critical determinants of performance and that comprehensive benchmarks are necessary to surface these effects [6].

Operational, environmental, and ethical considerations of AI for climate modeling

The deployment of machine learning (ML) and generative AI in climate science raises multiple operational and environmental concerns because training and running large models demands substantial data center resources, specialized hardware, extensive electricity consumption, and water for cooling, all of which contribute to the net environmental footprint of these tools [1,3,4]. At the same time, ML approaches offer trade-offs: surrogate models, emulators, and statistical downscalers can reduce the number of computationally expensive Earth system model (ESM) runs required, thereby lowering high-performance computing (HPC) energy use, but the energy and materials required to train and operate large ML models must be included in lifecycle assessments to determine net benefits [1,3,4]. Governance and reproducibility present parallel challenges, as the provenance and licensing of datasets (for example, CMIP6/Input4MIPs usage policies) and heterogeneity across climate model outputs complicate legal reuse and the ability to reproduce ML-based results, which is critical when informing policy decisions [1,4].

Managing uncertainty is central to trustworthy ML integration in climate science: combining generative approaches that can capture aleatoric or stochastic variability with multi-model ensembles that address epistemic or structural model uncertainty helps to distinguish sources of error and supports transparent benchmarking and evaluation [1,3]. Finally, the longer-term rise of powerful large language models (LLMs) and potential advanced AI capabilities suggests both opportunities and risks for decision support in climate policy; responsible integration will require transparent documentation, rigorous evaluation, and sustained human oversight to avoid over-reliance on opaque systems [2,1,3].

Environmental footprint and resource considerations

Generative AI and large ML architectures incur substantial operational resource demands during both training and inference phases: they require significant electricity for computation, water for data center cooling, and entail material- and energy-intensive chip manufacturing, all of which should be considered in net-benefit analyses of ML-enabled climate tools [1,3,4]. These resource demands can offset the computational savings that ML surrogates provide unless lifecycle accounting is performed and reported consistently [1,3,4].

Practical mitigation strategies focus on improving model efficiency and reuse. Techniques such as model distillation and pruning, transfer learning from pre-trained networks, and the reuse of trained emulators across multiple scenarios can materially reduce energy and resource requirements relative to training models from scratch for each task [1,3,4]. Complementing technical measures, transparent reporting of energy consumed during training and carbon accounting for published models enables more accurate assessments of environmental footprint and supports informed decision-making about when and how to deploy ML solutions in climate science [1,3,4].

Ethics, governance and reproducibility

Key governance challenges stem from the need for transparency about training data and model behavior when ML outputs are used to inform policy, since opaque data provenance or undisclosed preprocessing choices can undermine trust and lead to irreproducible conclusions [1,4]. Ensuring reproducibility requires standardized documentation of data sources, preprocessing pipelines, model architectures, and evaluation protocols so that results can be replicated across heterogeneous climate model outputs and different research groups [1,4].

Intellectual property and data licensing constraints also require careful management to enable legal and ethical reuse of both input climate datasets and ancillary corpora used for model training; compliance with dataset-specific terms (for example, those associated with multi-model archives) must be part of model development and dissemination workflows to avoid misuse and to preserve open scientific collaboration [1,4]. Clear communication of uncertainties and the limits of ML-derived inferences to stakeholders and policymakers is essential to prevent misinterpretation of model outputs in decision contexts [1,4].

Role and limitations of LLMs/AGI in climate modeling

LLMs and multimodal generative models can provide practical assistance across climate-modeling workflows, including extracting information from reports, automating preprocessing pipelines, synthesizing literature, and generating scenario narratives that improve accessibility and workflow efficiency; however, these capabilities are complementary to—and not replacements for—physics-based ESMs that embed mechanistic understanding of the climate system [2,1,3]. When used appropriately, language models can accelerate tasks that improve model provision and interpretation, but their outputs must be grounded in domain knowledge and validated against physical models and observations [2,1,3].

Claims about the arrival of artificial general intelligence remain contested; current LLMs exhibit strong language capabilities useful for knowledge extraction and decision support but have limitations in physical reasoning and require explicit grounding in physical models and datasets to be reliable in climate contexts [2,1,3]. Consequently, responsible integration of these models into climate policymaking and scientific workflows demands careful validation, transparency about model limitations, and preservation of human oversight to avoid undue reliance on opaque or insufficiently validated generative systems [2,1,3].

Conclusion

Applying AI to climate modeling advances capabilities in emulation, downscaling, and automated information extraction while posing practical and ethical challenges. Multi-model, ML-ready datasets like ClimateSet reduce barriers for large-scale ML research and enable training of emulators that capture inter-model uncertainty important for policy. Conditional generative models—especially diffusion-based approaches—offer robust tools for high-resolution downscaling and stochastic realization generation, outperforming conventional regression and addressing limitations of GANs in stability and fidelity. NLP tools built from large semi-structured questionnaires can automate extraction and structuring of climate disclosures, facilitating policy analysis and reducing manual effort. However, practitioners must account for the energy and resource costs of training large models, ensure reproducibility and transparency in preprocessing and model design, respect data licensing, and clearly communicate model uncertainty to decision-makers. Combining physics-based models with ML surrogates and ensemble strategies provides a pragmatic path: leverage AI for speed and pattern learning while maintaining physical grounding, uncertainty quantification, and human oversight to inform robust climate policy and research.

References

[1] Wikipedia contributors, "Generative artificial intelligence", Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Generative_artificial_intelligence

[2] Wikipedia contributors, "Artificial general intelligence", Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Artificial_general_intelligence

[3] Wikipedia contributors, "Applications of artificial intelligence", Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Applications_of_artificial_intelligence

[4] J. Kaltenborn, C. E. E. Lange, V. Ramesh, P. Brouillard, Y. Gurwicz, C. Nagda, J. Runge, P. Nowack, and D. Rolnick, "ClimateSet: A Large-Scale Climate Model Dataset for Machine Learning", arXiv, [Online]. Available: https://arxiv.org/abs/2311.03721

[5] N. Shidqi, C. Jeong, S. Park, E. Zeller, A. B. Nellikkattil, and K. Singh, "Generating High-Resolution Regional Precipitation Using Conditional Diffusion Model", arXiv, [Online]. Available: https://arxiv.org/abs/2312.07112

[6] D. Spokoyny, T. Laud, T. Corringham, and T. Berg-Kirkpatrick, "Towards Answering Climate Questionnaires from Unstructured Climate Reports", arXiv, [Online]. Available: https://arxiv.org/abs/2301.04253

Appendix A: Key points of Report

1. **Large-scale climate datasets and pipelines for ML (ClimateSet):**

   - ClimateSet provides a consistent, ML-ready, multi–climate-model dataset and modular pipeline that retrieves and preprocesses climate model inputs (Input4MIPs) and outputs (CMIP6) to support large-scale ML tasks in climate science.

   - Core ClimateSet composition: outputs from 36 Earth System Models (ESMs), input emission fields for multiple Shared Socioeconomic Pathway (SSP) scenarios (including historical), and a focus initially on key variables (temperature and precipitation); the core dataset is publicly available and extensible via the provided downloader and preprocessor.

   - Primary goals: (1) supply sufficient training data to enable large ML climate models (super emulators) that generalize across models and scenarios; (2) capture multi-model projection uncertainty important for policymaking, which single-model datasets cannot provide.

   - Design choices address common data challenges: harmonization of variable names/units, spatial/temporal alignment across models, and automated regridding to create consistent input/output tensors suitable for ML training and benchmarking.

   - ClimateSet demonstrates use as a benchmark for climate model emulation (temperature and precipitation), showing that ML models trained across multiple climate models gain insights into performance and generalization not obtainable from single-model datasets.

   - Limitations noted: remaining inconsistencies across models (e.g., ensemble members, vertical levels), selection of variables and scenarios in the core release, and dependence on CMIP6/Input4MIPs availability and ESGF infrastructure.

   - Core release aggregates 36 climate models' outputs and associated emission inputs, focusing initially on monthly variables like surface temperature and precipitation across several SSP scenarios and historical runs.

   - The dataset targets ML readiness: aligned grids, consistent units, and a standardized temporal resolution to enable direct model training without bespoke climate-domain preprocessing for each model.

   - ClimateSet provides a downloader that automates retrieval from ESGF and a preprocessor (including regridding and temporal aggregation) to create harmonized ML tensors; both components are modular so users can add variables, height levels, ensemble members, or additional scenarios.

   - The pipeline is intended to reduce domain-knowledge barriers for ML researchers and to enable reproducible dataset construction from CMIP6/Input4MIPs.

   - Primary ML use cases include climate model emulation (fast surrogates for long simulations), downscaling, scenario projection, extreme-event prediction, and training large multimodel emulators ('super emulators') that can quickly generate new scenario projections.

- Multi-model training facilitates quantification of model-projection uncertainty, a critical input for policy decisions and risk assessment, and provides larger training sample sizes aligned with ML best practices.

2. **Generative ML methods for climate downscaling and emulation:**

    - Generative models (GANs, VAEs, diffusion models, and conditional transformers) are applied to climate tasks such as downscaling (super-resolution), emulation (surrogate modeling), and stochastic weather/climate realization generation.
    - Diffusion-based generative models (denoising diffusion probabilistic models, DDPMs) conditioned on low-resolution climate variables have demonstrated superior fidelity for high upscaling factors (4x, 8x) in regional precipitation downscaling compared to regression baselines and are competitive with or advantageous over GANs in stability and preservation of high-frequency details.
    - Key architectural choices for effective downscaling include U-Net backbones for denoising networks, concatenation of upsampled LR conditioning variables with HR stochastic states, and loss formulations that predict injected noise (epsilon) using L2 objectives as in DDPM frameworks.
    - Conditional generative approaches allow sampling multiple plausible high-resolution realizations from the same low-resolution conditioning, enabling uncertainty quantification at local scales—a major advantage over deterministic regression downscalers.
    - Emulation benefits from large, multi-model datasets (e.g., ClimateSet), allowing ML surrogates to learn cross-model response patterns and generalize across forcing scenarios, drastically reducing computational cost compared to full ESM runs.
    - Method: adapt DDPM to map low-resolution (LR) climate inputs X_LR to high-resolution (HR) precipitation X_HR using a reverse diffusion chain p_theta(X_HR^{t-1} | X_HR^t, X_LR) modeled as Gaussians with learned mean mu_theta and variance sigma_t^2.
    - Training objective: predict injected Gaussian noise epsilon via a neural denoiser epsilon_theta (U-Net), optimizing L = E[||epsilon - epsilon_theta(X_LR, X_HR^t, t)||^2], with forward process q defined by scheduled beta_t noise variances.
    - Implementation details: condition on multiple LR variables (e.g., TS, PRECT, dPHIS), upsample LR inputs to HR grid for concatenation, and use iterative sampling (T steps) to generate HR realizations; shown effective on CESM simulation data for North America.
    - Comparisons: diffusion models outperform regression-based super-resolution and often yield more stable training than GANs, particularly at high scaling factors where regression tends to oversmooth and GANs can be unstable to train.
    - Evaluation metrics should include pixel-wise skill (RMSE, MAE), distributional statistics (e.g., precipitation frequency/intensity distributions), spatial structure metrics, and out-of-sample generalization across time slices or forcing scenarios to assess robustness.

- Practical considerations: selecting conditioning variables, handling temporal consistency, and computational cost of iterative sampling are trade-offs; ensembles of stochastic samples can quantify aleatoric uncertainty while training across multiple climate models captures epistemic uncertainty.

3. **Natural language methods for climate information extraction and benchmarking:**

   - Large volumes of unstructured climate reports (corporate, city, state, national disclosures) require NLP tools to convert free-form text into structured questionnaires and fields useful for policy analysis, monitoring, and research.
   - The CLIMA work introduces two large questionnaire-derived datasets (CLIMA-CDP drawn from the Carbon Disclosure Project; CLIMA-INS from insurance disclosures) and CLIMABENCH, a benchmark aggregating multiple climate text classification tasks to evaluate models on climate-specific NLP challenges.
   - Approach: utilize existing semi-structured questionnaires to self-supervise classifiers that align text passages to questionnaire questions, enabling in-domain training without costly manual annotation; experiments show models can generalize across stakeholder types (cities, corporations, states).
   - A counterintuitive empirical finding: general-purpose language models often outperform domain-specific models (e.g., ClimateBERT) on compiled benchmark tasks, indicating model choice and pretraining strategy impact domain adaptation.
   - Applications include automating questionnaire population, accelerating evidence aggregation for policy, detecting greenwashing, and building datasets to fine-tune LLMs for climate policy tasks.
   - CLIMA-CDP: derived from CDP disclosure questionnaires with broad coverage (thousands of organizations, many questions); CLIMA-INS: NAIC insurance climate risk survey responses yielding ~17k question-answer pairs after preprocessing.
   - CLIMABENCH aggregates multiple climate NLP datasets to provide standardized evaluation across tasks such as topic classification, question-answer alignment, and report-to-question mapping.
   - Self-supervised classification using questionnaire structure enables training without manual annotation; models were tested on in-domain, cross-domain, and real-world mapping tasks (unstructured reports to questionnaires).
   - Empirical results indicate that domain-tailored pretraining is not always superior; well-tuned general models can outperform specialized domain models on downstream climate document tasks, underscoring the importance of evaluation on comprehensive benchmarks.

4. **Operational, environmental, and ethical considerations of AI for climate modeling:**

   - Computational cost and environmental footprint: training and deploying large ML models (including generative models and LLMs) require extensive data center resources, specialized hardware, significant electricity consumption, and water for cooling—raising concerns about the net environmental impact of AI development and use in climate science.

- Trade-offs: while ML surrogates and downscalers can reduce the need for computationally expensive ESM runs (thus saving HPC energy), the energy required to train large ML models and run inference at scale must be accounted for in lifecycle assessments.

- Governance, reproducibility and IP: issues include dataset provenance and licensing (e.g., CMIP6/Input4MIPs usage policies), reproducibility across heterogeneous climate model outputs, model interpretability for policy decisions, and intellectual property or copyright concerns when using publicly sourced data in generative systems.

- Uncertainty management: combining ML approaches with multi-model ensembles (as facilitated by ClimateSet) is important for separating aleatoric/stochastic uncertainty (addressable via generative sampling) and epistemic/model structural uncertainty (addressable by multi-model training and transparent benchmarking).

- Longer-term considerations: the rise of powerful LLMs and potential AGI-like capabilities may influence decision-support tools in climate policy, but responsible integration requires transparency, evaluation, and human oversight to avoid over-reliance on opaque models.

- Generative AI and large ML models consume substantial electricity and cooling water during training and inference; data center footprint and chip manufacturing material intensity must be considered in net-benefit analyses of ML-enabled climate science tools.

- Practical mitigation strategies include model efficiency (distillation, pruning), transfer learning from pre-trained models, reuse of emulators across scenarios, and reporting energy-of-training and carbon accounting for published models.

- Key governance issues involve transparency of training data and model behavior when informing policy, reproducibility of ML-based climate results across preprocessing choices, and clear communication of uncertainties to stakeholders and policymakers.

- Intellectual property and data licensing constraints (for both input climate data and ancillary training corpora) must be managed to ensure legal and ethical reuse.

- LLMs and multimodal generative models can assist climate modeling workflows (data extraction from reports, automating preprocessing pipelines, literature synthesis, scenario narrative generation) but are not substitutes for physics-based ESMs; they can complement domain models for speed and accessibility.

- Claims about AGI remain contested; while current LLMs provide powerful language capabilities (useful for knowledge extraction and decision support), their limitations in physical reasoning and requirement for grounding in physical models mean they should be used with careful validation in climate contexts.

Appendix B: Recent News

- **Artificial intelligence for modeling and understanding extreme weather and climate events - Nature**
  - Nature - Published on Mon, 24 Feb 2025 08:00:00 GMT
  - For more details click here.
- **Artificial intelligence is helping improve climate models - The Economist**
  - The Economist - Published on Wed, 13 Nov 2024 08:00:00 GMT
  - For more details click here.
- **Accelerating Climate Modeling with Generative AI - UC San Diego Today**
  - UC San Diego Today - Published on Mon, 02 Dec 2024 08:00:00 GMT
  - For more details click here.
- **Advancements and challenges of artificial intelligence in climate modeling for sustainable urban planning - Frontiers**
  - Frontiers - Published on Mon, 19 May 2025 07:00:00 GMT
  - For more details click here.
- **Leveraging artificial intelligence for research and action on climate change: opportunities, challenges, and future directions - ScienceDirect.com**
  - ScienceDirect.com - Published on Tue, 01 Jul 2025 07:00:00 GMT
  - For more details click here.
- **How AI could shape the future of climate science - American Physical Society**
  - American Physical Society - Published on Mon, 16 Jun 2025 07:00:00 GMT
  - For more details click here.
- **At the Intersection of Climate and AI, Machine Learning is Revolutionizing Climate Science - Georgia Institute of Technology**
  - Georgia Institute of Technology - Published on Wed, 22 Jan 2025 08:00:00 GMT
  - For more details click here.
- **Charting the Intersection of Climate Change and AI - The Regulatory Review**
  - The Regulatory Review - Published on Tue, 24 Dec 2024 08:00:00 GMT
  - For more details click here.
- **AI: Good for Climate Models; Bad for Climate - Astrobites**
  - Astrobites - Published on Thu, 12 Jun 2025 07:00:00 GMT
  - For more details click here.
- **How AI is Making Climate Modeling Faster, Greener, and More Accurate | NVIDIA Technical Blog - NVIDIA Developer**
  - NVIDIA Developer - Published on Wed, 04 Dec 2024 08:00:00 GMT

- For more details click here.

- **Simpler models can outperform deep learning at climate prediction - MIT News**
  - MIT News - Published on Tue, 26 Aug 2025 07:00:00 GMT
  - For more details click here.

- **AI-Based Climate Modelling Market Size | CAGR of 24% - Market.us**
  - Market.us - Published on Tue, 07 Jan 2025 06:17:10 GMT
  - For more details click here.

- **IBM and NASA Release Open-Source AI Model on Hugging Face for Weather and Climate Applications - IBM Newsroom**
  - IBM Newsroom - Published on Mon, 23 Sep 2024 07:00:00 GMT
  - For more details click here.

- **Zooming in: Efficient regional environmental risk assessment with generative AI - Google Research**
  - Google Research - Published on Thu, 05 Jun 2025 07:00:00 GMT
  - For more details click here.

- **AI for Climate Change: Opportunities and Challenges - eWeek**
  - eWeek - Published on Tue, 15 Oct 2024 07:00:00 GMT
  - For more details click here.

- **AI-empowered next-generation multiscale climate modelling for mitigation and adaptation - Nature**
  - Nature - Published on Wed, 25 Sep 2024 07:00:00 GMT
  - For more details click here.

- **NOAA Research develops an AI-powered sibling to its flagship weather model - National Oceanic and Atmospheric Administration (.gov)**
  - National Oceanic and Atmospheric Administration (.gov) - Published on Tue, 22 Jul 2025 07:00:00 GMT
  - For more details click here.

- **New NVIDIA Earth-2 Generative AI Foundation Model Simulates Global Climate at Kilometer-Scale Resolution - NVIDIA Blog**
  - NVIDIA Blog - Published on Tue, 10 Jun 2025 07:00:00 GMT
  - For more details click here.

- **Supercomputers and AI are helping build better climate models - The Economist**
  - The Economist - Published on Wed, 20 Nov 2024 08:00:00 GMT
  - For more details click here.

- **Enhancing system resilience to climate change through artificial intelligence: a systematic literature review - Frontiers**

- Frontiers - Published on Mon, 30 Jun 2025 08:35:04 GMT
- For more details click here.