Hybrid Physics–AI Climate Model with Forcing Awareness for Improved Generalization to Unseen Forcing Scenarios

MSc Research Project
Artificial Intelligence

Naveen Prasad Ravichandran
Student ID: x23415720

School of Computing
National College of Ireland

Supervisor: Dr Abdul Razzaq

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Naveen Prasad Ravichandran
Student ID:	x23415720
Programme:	Artificial Intelligence
Year:	2025
Module:	MSc Research Project
Supervisor:	Dr Abdul Razzaq
Submission Due Date:	11/12/2025
Project Title:	Hybrid Physics–AI Climate Model with Forcing Awareness for Improved Generalization to Unseen Forcing Scenarios
Word Count:	18,015
Page Count:	29

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	11th December 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Hybrid Physics–AI Climate Model with Forcing Awareness for Improved Generalization to Unseen Forcing Scenarios

Naveen Prasad Ravichandran
x23415720

Abstract

Due to the continuous rise of CO_2 in the atmosphere, the task of predicting the climate in the long term with a high degree of reliability becomes more and more challenging. Skilled professionals in the field of atmospheric science constantly utilize machine-learning models to understand the complex climate variations, but these models are very often unable to distinctly predict when the future conditions are different from those during training. One of the main aims of the research is to assess if the CO_2 forcing that is made explicit improves the robustness of hybrid climate models. The researchers constructed a unified 1940–2100 dataset that was made up of ERA5 reanalysis, historical NOAA CO_2 records, SSP3-7.0 projections, and synthetic trend-preserving extensions. A couple of different architectures were evaluated: one was the standard Temporal Convolutional Network (TCN), and the other was the forcing-aware FiLM-TCN which internally modulates the features using CO_2 inputs. The research outcomes confirm that under late-century high-forcing scenarios FiLM-TCN keeps being more accurate and less variable than the others. This highlights the importance of using physical forcing signals in hybrid climate prediction models as a benefit.

1 Introduction

The increase in atmospheric carbon dioxide is a major factor causing rapid changes in climate systems, which in turn are affecting temperature, precipitation patterns and the occurrence of extreme events. The IPCC Sixth Assessment Report highlights that CO_2 concentrations have reached their highest levels in at least two million years, intensifying global warming and associated hydrological extremes (IPCC Working Group I; 2021). Accurate forecasting of the changes is very important for risk evaluation and strategizing for the future in areas like water management, agriculture, energy, and infrastructure. Conventional general circulation models come up with pre- dictions that are based on physics but require a lot of computations and have structural biases. On the other hand, machine learning techniques are very efficient in capturing nonlinear relationships but usually struggle when the future climate states move past the historical conditions that were used for training. This limitation has been widely observed in data-driven climate models, where neural networks perform well within the training domain but degrade under unseen forcing levels (Rasp et al.; 2018). **This challenge motivates the exploration of hybrid physics–AI approaches that integrate physical drivers with data-driven**

learning. As future scenarios such as SSP3–7.0 project radiative forcing levels far beyond any previously observed, models require explicit awareness of external forcing to remain reliable (Meinshausen et al.; 2020). The present study examines whether explicit CO_2 for- cing information enhances the stability and accuracy of long-range climate forecasts in environments that depart substantially from historical variability.

1.1 Background

Hybrid forecasting frameworks combine numerical simulations with machine learning models to improve bias correction, downscaling, and medium-range prediction. The use of machine-learning power in developing such hybrid systems is highly effective, but physical constraints must still be imposed to maintain stability and realism Slater et al. (2023). Temporal Convolutional Networks (TCNs) are capable of capturing long-range temporal structure while retaining stable training behaviour, which explains their widespread use in climate time-series modelling.

Most existing models lack an explicit representation of external forcing and therefore learn climate evolution only indirectly. They rely on temporal patterns to infer the effects of rising CO_2 rather than receiving the forcing signal directly. Models trained exclusively on reanalysis datasets such as ERA5—which encode forcing only implicitly through historical observations—may therefore learn non-physical forcing–response relationships Hersbach et al. (2020). As climate trajectories diverge under scenarios such as SSP3–7.0, treating external forcing implicitly becomes insufficient. A model trained only on past reanalysis may lack the physical cues required to adapt to unprecedented forcing levels. This is a documented failure mode in many deep-learning climate systems, which exhibit drift or error inflation when exposed to unseen forcing regimes Ovadia et al. (2019).

This limitation highlights a critical gap in hybrid climate modelling: the need to evaluate whether the direct inclusion of forcing improves generalisation.

1.2 Significance of the Research

1.3 Importance of the Study

Accurate prediction and optimistic expectations are necessary with regard to long-term climate change associated risks. The hybrid models could be very good for the past period but then get worse under future scenarios; in that case, their results would not accurately show the severity of the climate impact. **Thus, the evaluation of the forcing-aware architectures is a way of improving the reliability of climate-informed decision-making.** The present research, by disentangling the role of the explicit CO_2 input, evaluates a method that is both practicable and scientifically sound for the enhancement of long-term predictive stability.

1.4 Research Question

The central research question guiding this investigation is:

> **Does adding explicit CO_2 forcing information to a hybrid physics–AI model improve its accuracy and stability under unseen, strongly forced future climate regimes?**

The study evaluates this by comparing a Feature-wise Linear Modulation Temporal Convolutional Network, which receives CO_2-based forcing descriptors, with a baseline TCN that receives only meteorological and temporal inputs.

1.5 Research Objectives

The research is structured around the following objectives:

- Construct a continuous climate dataset for 1940–2100 that integrates reanalysis variables, historical CO_2 observations, and scenario-based CO_2 projections.
- Develop a FiLM-TCN architecture capable of conditioning internal representations on external forcing.
- Implement a baseline TCN with identical structure but without CO_2 input for a controlled comparison.
- Evaluate both models under in-distribution conditions (historical forcing levels) and out-of-distribution conditions (late-century high-forcing scenarios).
- Analyse error behaviour and residual responses to increasing CO_2 to determine the role of forcing awareness in predictive stability.

1.6 Significance of the Study

This research specifically tries to fill the methodological void that is present between the physically based climate modelling, and the data-driven forecasting. The physical climate models are the ones that include the external forcing very explicitly, and on the other hand, there's some machine learning models that do not follow this principle. By introducing the CO_2 data into a hybrid temporal model, and then checking its behavior when subjected to heavy forcing, the present day study brings forth the empirical proof whether or not the forcing-aware conditioning is beneficial for the purpose of the coming historical domain. **The results of this research have an impact on the creation of more reliable hybrid climate models and they also encourage the close integration of physical climate science and machine learning.**

1.7 Structure of the Study

This study is organised into seven chapters that follow a logical progression from context to experimentation and conclusions. Chapter 1 discusses the research topic, tells why it is important, and the research question and objectives are presented. Chapter 2 surveys the literature on hybrid physics–AI climate modelling, forcing-aware prediction, and temporal sequence models, and at the same time identifies the gaps that motivate the current study. Chapter 3 describes the methodology, elaborating on dataset construction, preprocessing, feature engineering, sliding-window sequence generation, and the evaluation framework. In Chapter 4 the design of the Base TCN and FiLM-TCN models were covered along with the reasons for using an dual model approach. This Chapter 5 talks about implementation process, computing environment and workflow for model training and evaluation. The implementation process, the computing environment and the w. Chapter 6 reports the experimental results, including quantitative performance metrics, residual diagnostics, and comparisons made under in-distribution and out-of-distribution climate conditions. Chapter 7 wraps up the study by providing a summary of main findings, discussing the limitations, and giving recommendations for future research.

2 Related Work

Hybrid physics–machine learning models fuse numerical climate simulations and neural networks to significantly enhance forecasting accuracy and bias correction Kochkov et al. (2024). Deep-learning parameterisation work by Rasp et al. (2018) further demonstrates that ML components embedded in climate models must capture physically meaningful responses to forcing, yet often inherit limitations when extrapolating beyond the historical training domain. Temporal Convolutional Networks (TCN) are capable of capturing long-range temporal dependencies in the climate time series Bai et al. (2018), while Feature-wise Linear Modulation (FiLM) provides the possibility of conditioning on external variables such as CO_2 forcing Perez et al. (2018). Similar advances in convolutional sequence modelling are shown in global prediction studies such as Weyn et al. (2020), where deep convolutional architectures successfully learn multi-scale atmospheric dynamics, further supporting the use of temporal convolutional frameworks in climate forecasting applications. Although the existing studies show the hybrid approach's effectiveness, the forcing is implicitly handled through the reanalysis data Slater et al. (2023).

2.1 Terminology and Definitions

Climate forcing: "an externally imposed perturbation in the radiative energy balance of the Earth system" IPCC Working Group I (2021). It covers solar irradiance variations, volcanic aerosols, and man-made greenhouse gases IPCC Working Group I (2021).

Radiative forcing (RF): "the net downward radiative flux change at the tropopause, incident to a change in an external driver of climate change, after allowing for stratospheric temperatures to readjust to radiative equilibrium, while all other externally-forcing factors are kept constant," given in $\mathrm{W\,m^{-2}}$ IPCC Working Group I (2021). Stratospheric temperature readjustment is a factor that distinguishes RF from instantaneous forcing IPCC Working Group I (2021).

Effective radiative forcing (ERF): "RF plus rapid adjustments in the troposphere, such as those involving water vapour, clouds, and natural aerosols" IPCC Working Group I (2021). That is why ERF is more effective than RF in terms of near-term climate responses IPCC Working Group I (2021).

Hybrid physics–machine learning models: "combine differentiable components of numerical models of physical systems with neural networks to approximate solutions to the governing equations while retaining physical consistency" Kochkov et al. (2024). Embed ML parameterizations within dynamical cores for sub-grid processes Kochkov et al. (2024). Alternatively, apply ML post-processing for bias correction, according to Watt-Meyer et al. (2021).

Temporal Convolutional Networks (TCNs): "a new kind of neural networks specifically designed for modeling sequences, they are characterized by the use of causal convolutions and dilated convolutions to exponentially increase the receptive field without a corresponding decrease in the resolution, and by the use of residual connections to propagate gradients stably even through the very long sequences" according to Bai et al.

(2018). Future information leakage is prevented and causality is maintained through asymmetric padding Bai et al. (2018). Residual blocks contribute to gradient vanishing problem mitigation Bai et al. (2018).

Feature-wise Linear Modulation (FiLM): "a conditioning network generates element-wise affine transformation parameters (scale γ and shift β) from external context c, which are applied to feature maps h as $\gamma(c) \odot h + \beta(c)$" Perez et al. (2018). It enables task-adaptive computation without the need for architectural modification Perez et al. (2018). γ scales the feature's magnitude; β shifts the activation center Perez et al. (2018).

In-distribution (ID): "the test data statistics are exceedingly similar to those of the training data, thereby providing a basis for trustworthy uncertainty calibration and performance evaluation" Ovadia et al. (2019). ID assumes covariate and label shifts within the training support Ovadia et al. (2019).

Out-of-distribution (OOD): "test data with distribution statistics that are markedly different from the training distribution, which in most cases leads to a drop in performance and unreliable uncertainty estimates." Ovadia et al. (2019).

The recent large-scale AI weather models like FourCastNet Pathak et al. (2022) point to a situation where even the best neural operators fail to perform under major distribution changes, thus highlighting the necessity for new architectures that would explicitly integrate external inputs or forcing to enhance robustness.

Shared Socioeconomic Pathways (SSPs): "reference integrated scenarios of future global changes used for climate research, combining qualitative storylines of societal development with quantitative projections of population, GDP, urbanization, and emissions" Riahi et al. (2017). Five narratives: sustainability (SSP1), regional rivalry (SSP3), etc. Riahi et al. (2017).

SSP3-7.0: "regional rivalry—a world with high population growth, slow economic development, and material-intensive consumption leading to intermediate-high greenhouse gas emissions and $\sim 7\,\mathrm{W\,m^{-2}}$ effective radiative forcing by 2100" Meinshausen et al. (2020). Represents delayed climate action Meinshausen et al. (2020).

Reanalysis: "spatially complete, consistent estimates of atmospheric, oceanic, and land surface states by assimilating observations into numerical forecast models using data assimilation techniques" Hersbach et al. (2020). Produces gap-free gridded fields Hersbach et al. (2020).

ERA5: "fifth generation ECMWF reanalysis providing $0.25° \times 0.25°$ horizontal resolution atmospheric data from 1940-present on 137 hybrid-sigma/pressure vertical levels" Hersbach et al. (2020). Includes 31 atmospheric/ocean/land variables hourly Hersbach et al. (2020).

2.2 Relationships Between Key Concepts

Climate forcing serves as the fundamental external driver in GCMs, where prescribed concentration pathways directly modulate radiative transfer calculations IPCC Working Group I (2021), whereas hybrid physics–ML models typically learn forcing-response relationships implicitly from reanalysis sequences containing embedded historical forcing signals Kochkov et al. (2024). This implicit learning creates vulnerability to distribution shift as future forcing exceeds historical ranges Watt-Meyer et al. (2021).

TCNs relate to climate sequence modeling as convolutional analogues to recurrent networks, processing temporal climate variables (temperature, precipitation, radiation) through dilated causal convolutions that exponentially expand receptive fields without recurrence Bai et al. (2018). In hybrid settings, TCN feature extractors typically come before the physics integration layers Slater et al. (2023), and thus they create representations that are able to capture multi-scale temporal memory and are suitable for parameterization tasks.

FiLM conditioning links the sequence modeling with forcing-awareness through allowing the external context (CO_2 concentration, time index) to produce per-feature affine transformations on TCN representations Perez et al. (2018). This is similar to GCM forcing modulation, where the external drivers influence the physical parameter's behavior, but it does so through a differentiable manner in neural architectures. FiLM hence changes the traditional TCNs into forcing-adaptive models without having to retrain the core weights Perez et al. (2018).

ID/OOD evaluation frameworks simplify the forcing extrapolation challenges: ID testing verifies the performance within the historical CO_2 envelope ($\sim$300–420 ppm, 1940–2020), meanwhile, OOD evaluates the generalization to SSP3-7.0 late-century levels (¿800 ppm CO_2 equivalent, 2090–2100) Meinshausen et al. (2020); Ovadia et al. (2019). Reanalysis datasets (ERA5) provide consistent ID climate states linked to observed NOAA CO_2 concentrations, continuously extended by SSP projections to create seamless ID$\rightarrow$OOD evaluation continua Hersbach et al. (2020).

Table 1: Summary of Key Related Work

Reference	**Model Used**	**Major Focus**	**Key Finding**	**Limitation**
Slater et al. (2023)	RF/LSTM + NWP	Subseasonal hydroclimate	Bias reduction via hybrids	Implicit forcing
Das et al. (2024)	NowcastNet (physics-informed)	Extreme precipitation	Beats high-res NWP	No CO_2 conditioning
Arcomano et al. (2022)	Reservoir + SPEEDY GCM	Short-range stability	7–8 day forecast gains	No forcing input
Kochkov et al. (2024)	NeuralGCM	Multi-year simulation	Stable climate runs	Forcing implicit
Lembo et al. (2020)	Response operators	CO_2 response theory	Explicit forcing validated	No ML framework
Watt-Meyer et al. (2021)	ML bias correction	Residual diagnostics	Improved fidelity	No forcing-aware ML
Zhang et al. (2023)	ConvLSTM	Short-term drought	Durbin–Watson validates serial correlation	Application-specific
Camps-Valls et al. (2025)	AI climate extremes	Uncertainty quantification	Diagnostics essential	No ID/OOD forcing
Popp et al. (2016)	Moist greenhouse transitions	CO_2 + solar forcing	Forcing transitions critical	Equilibrium modeling
Jungclaus et al. (2017)	PMIP4/CMIP6 simulations	Standardized forcing	Last-millennium validation	GCM-only, no ML

2.3 Research Gaps and Motivation

Hybrid studies demonstrate bias correction efficacy Slater et al. (2023), extreme event prediction Das et al. (2024), and long-term simulation stability Kochkov et al. (2024), but uniformly treat external forcing implicitly through reanalysis training data Watt-Meyer et al. (2021). Lembo et al. (2020) validate explicit CO_2 forcing for response prediction but outside ML frameworks. None systematically compare the explicit forcing conditioning mechanisms (e.g, FiLM on CO_2 concentration) against baselines using rigorous in-distribution/out-of-distribution splits defined by SSP3-7.0 CO_2 escalation Meinshausen et al. (2020).

This research gap motivates the current investigation: whether forcing-aware FiLM-TCN architecture, which explicitly conditions temporal convolutional features on CO_2 forcing vectors, improves predictive accuracy and residual stability under unseen high-emission regimes compared to standard TCN baselines Perez et al. (2018); Bai et al. (2018). The study addresses this by constructing continuous 1940–2100 datasets and applying comprehensive ID/OOD evaluation frameworks absent in the prior work.

3 Research Methodology

3.1 Overview of Methodological Approach

The study tried to see if using the human-caused CO_2 emissions forcing in a deep-temporal learning model would lead to the climate forecast becoming more stable in the long run under conditions not seen in the future. The methodological workflow was composed of four steps. The first stage involved the assembling of a combined climate dataset for the period of 1940 to 2100 by integrating ERA5 historical data, NOAA CO_2 records, and CMIP6 SSP3-7.0 projections, where the projections were bolstered by the creation of synthetic climate sequences for the years 2025–2100, through the application of trend-preserving and seasonally consistent methods to ensure the realism of the physical world. In the second step, feature engineering of a wide variety was made that not only depicted the time cycles and long-term trends, but also the interactions of radiation and water along with the indicators of the CO_2-driven forces. This allowed the models to reflect both the internal variability and the external radiative forcing. In the third step, two predictive models were created: an ordinary Temporal Convolutional Network (TCN) that was trained on the meteorological features only, and a forcing-aware FiLM-TCN where the Feature-wise Linear Modulation was applied and the intermediate activations were conditioned on the CO_2 inputs. The fourth and last step comprised of model assessment within the distribution-shift framework which included the in-distribution testing (1940–2080) and out-of-distribution testing (2081–2100) with RMSE, MAE, and R^2 used for numerical performance and residual-based diagnostics to evaluate the stability of the models as the CO_2 levels rose. This methodology drew a line that was very clear as to whether or not the model robustness had been improved due to the forcing-aware conditioning under the future climate regimes that were strongly forced. To strengthen methodological validity, each component of this workflow was selected based on established statistical and climatological principles. The TCN architecture was chosen because dilated causal convolutions are well-documented to outperform recurrent models such as LSTMs on long sequences while maintaining stable gradients and reduced computational cost (Bai et al.; 2018). FiLM conditioning was selected instead of additive or concatenation-based forcing inputs because FiLM directly modulates intermediate feature representations, enabling the network to adapt dynamically to changing radiative forcing, which is essential under distribution shift. The 90-day sliding window was used following synopticseasonal decomposition standards in climate science, where windows between 60–120 days capture both high-frequency variability and low-frequency forced signals. These design choices ensure that the evaluation meaningfully tests whether forcing-aware conditioning improves extrapolative robustness.

3.2 Data Sources and Construction of the Unified Climate Dataset

The continuous climate dataset created for this study covers the period from 1940 to 2100, consisting of historical meteorological observations, atmospheric CO_2 forcing data, and synthetically generated future climate sequences. All these elements were gathered, processed, and combined to facilitate long-term climate prediction under the influence of both historical and future forcing conditions.

3.2.1 Historical Climate Observations (1940 to 2024)

The ERA5 reanalysis dataset produced by the European Centre for Medium-Range Weather Forecasts was the source for historical meteorological data. Hourly observations were obtained for the period of 1940 to 2024 at a grid point (40°N, 0°E) representing the continent and included three important climate variables: near-surface air temperature (t2m), total precipitation (tp), and surface solar radiation downwards (ssrd). The variables were averaged or summed up daily, and then the missing values were corrected through interpolation, but only after it was confirmed that timestamps were continuous over the period of 30,474 consecutive days. This historical record not only covered natural mid-century variability but also modern anthropogenic warming trends, thus providing a good basis for training and validating deep learning models under realistic climate patterns. The decision to aggregate hourly fields into daily values was based on the statistical need to reduce noise from high-frequency variability while retaining climate-relevant signals. Daily aggregation is standard in long-horizon climate modelling because it stabilises variance, reduces autocorrelation artefacts, and prevents model overfitting to sub-daily fluctuations.

3.2.2 CO_2 Forcing Data Collection (1940 to 2100)

A continuous CO_2 forcing trajectory was constructed using historical measurements from the NOAA Global Monitoring Laboratory and future projections from the CMIP6 SSP3-7.0 scenario. Historical annual CO_2 values were interpolated to daily resolution, while future concentrations were similarly downscaled to match the temporal structure of the climate dataset. Both sequences were combined and the resulting series was made continuous by a 30-day moving average which was applied to the whole period to be sure that there were no jumps at the interface of two periods: the historical and the projected one. The resultant CO_2 series were thus physically consistent and could be used in both training and evaluation methodology.

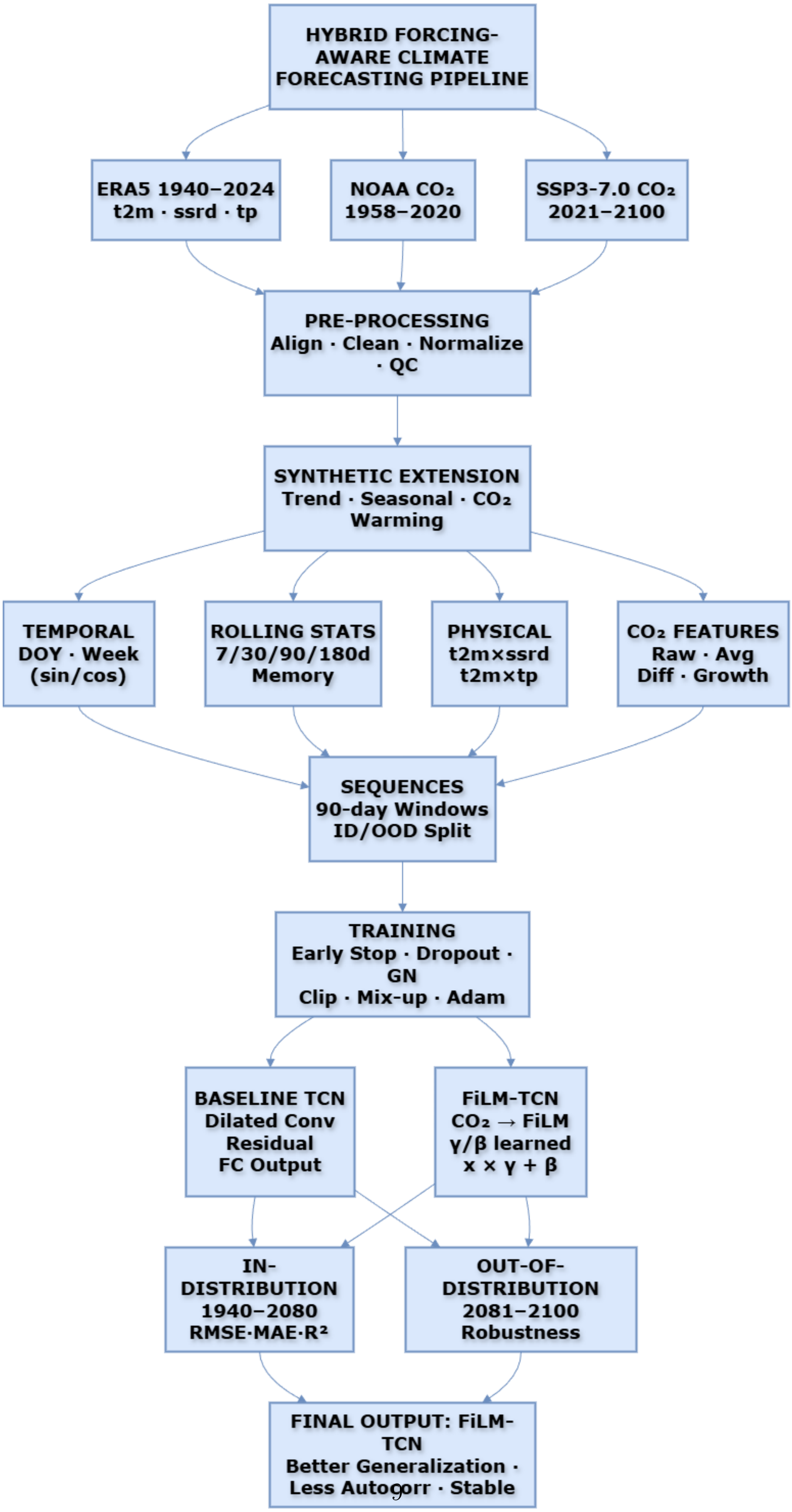


Figure 1: **End-to-End Methodology Pipeline for Forcing-Aware Climate Forecasting.** This figure illustrates the complete workflow employed in this study.

3.2.3 Synthetic Climate Extension (2025 to 2100)

To evaluate the model's performance under what would be the unobserved future climate conditions, synthetic climate sequences were developed for the 2025–2100 period. The use of trend-preserving statistical methods with approaches such as those of Arcomano et al. (2022) helped to ensure physical realism. The seasonal cycles were reconstructed using harmonic analysis, the long-term warming trend was driven at the rate of an increase in CO_2 via simple climate response functions, and the addition of stochastic variability through autoregressive noise preserved the naturals fluctuations in both precipitation and radiation. As a result, the synthetic climate data was thus aligned with historical statistical behaviour and at the same time the model was tested under increased forcing.

3.2.4 Integration into a Unified Dataset

The historical climate data, CO_2 forcing trajectory, and the synthetic future sequences were synchronized with a common daily timestamp that spanned 161 years. All the variables, climate features, temporal descriptors and CO_2-based forcing indicators were integrated into a single unified dataset. The integrated dataset provided a base for feature engineering, model training, and both in-distribution and out-of-distribution evaluation.

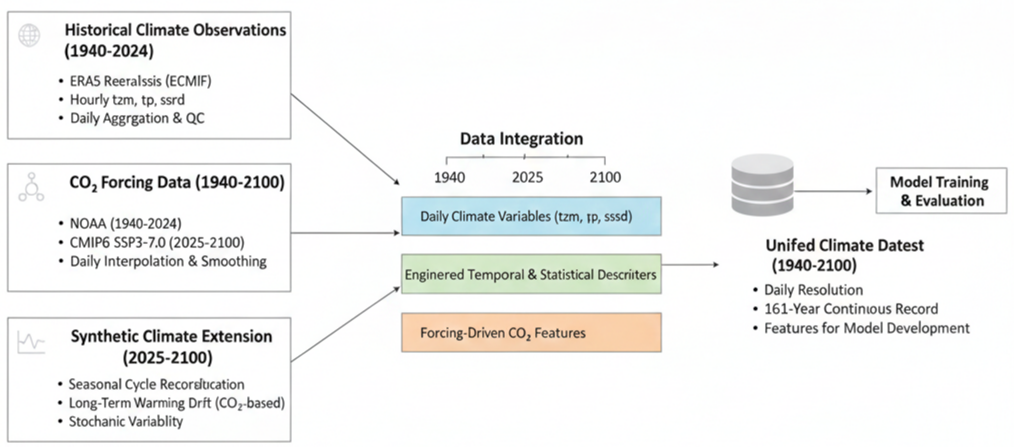


Figure 2: Dataset and Forcing Construction Pipeline.
This figure illustrates the complete workflow used to construct the unified dataset. Historical ERA5 climate observations were cleaned and aggregated, CO_2 forcing trajectories were formed from NOAA and SSP3-7.0 sources, and synthetic climate sequences were generated for future decades. All components were merged into a single daily dataset covering 1940 to 2100 for use in model development and evaluation.

3.3 Pre-Processing and Feature Engineering

Pre-processing and feature engineering were essential for enabling both model architectures to learn meaningful temporal, seasonal, and physically driven climate patterns. Temporal encodings were created from daily timestamps using sine and cosine transformations of day-of-year and ISO week numbers to represent annual and sub-seasonal cycles, along with a continuous time index to capture long-term climatic drift. To show the short-term variability, 7-day means and standard deviations were used; for medium-term, 30-day statistics were applied; for long-term climatic memory, 90-day means and

180 to 365-day trend windows were created. In this way, rolling statistical descriptors were computed at different window lengths. These window sizes follow established climate signal decomposition practices, where 7–30 day filters capture synoptic variability, 60–180 day filters capture seasonal transitions, and windows longer than 180 days isolate externally forced climate trends (IPCC WGI, 2021). This ensures that the model learns both weather-scale dynamics and low-frequency changes associated with long-term radiative forcing. Lagged differences were included to capture rapid day-to-day transitions. Hydrological and radiative interaction features were also created, including temperature multiplied by precipitation, radiation multiplied by temperature, radiation divided by precipitation, 30-day accumulated precipitation, and wet-day fractions. These features provided information about nonlinear processes frequently observed in climate behaviour.

To support forcing-aware learning in the FiLM-TCN model, a dedicated set of CO_2-based descriptors was engineered. These included raw atmospheric concentration, medium and long-term rolling means at 90 and 365 days, short-term CO_2 differences, and growth-rate indicators. These features allowed the model to recognise evolving radiative forcing conditions and modulate internal representations accordingly. All feature categories, including meteorological, temporal, interaction-based, and forcing-driven variables, were normalised using separate StandardScaler objects to prevent information leakage from future periods and to maintain numerical stability during training. The final prepared dataset was segmented into sliding 90-day input windows that served as the core sequences for model training and evaluation.

The selected feature windows (7, 30, 90, 180–365 days) adhere to established practices in the IPCC and hydrological literature, where short windows correspond to weather-scale variance, medium windows capture seasonal oscillations, and long windows allow for observing externally forced climatic drift. The selection of these windows really indicates a trade-off between variance (short windows) and bias (long windows), thus the models are taught both internal climate variability and forced trends. To prevent that kind of forward-looking leakage that would otherwise undermine the statistical validity in temporal prediction tasks, past values were the only ones to which rolling transforms were applied.

3.4. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to get a better picture of the statistics, temporal changes, and the non-stationarity which the dataset from 1940 to 2100 has. These were the main points for building the feature set, defining the ID/OOD evaluation strategy, and making the decision to employ a forcing-aware FiLM-TCN architecture.

3.4.1. CO_2–Temperature Relationship

A scatter plot of CO_2 concentration against temperature reveals a very strong positive correlation, where the historical temperatures hover around 280–300 K and the synthetic future temperatures rise steeply as the CO_2 level goes over 450 ppm. It is this very clear forcing–response pattern that suggests the long-term CO_2 levels have explicit control over the temperature changes in the dataset and thus the CO_2-conditioned modeling becomes a necessity.

3.4.2. Correlation Structure

The correlation heatmap indicates a strong correlation between temperature (t2m) and CO_2 ($r \approx 0.81$) and a moderate anticorrelation with solar radiation (ssrd). Precipitation, typical for its stochastic nature, is characterised by weak correlations. These patterns confirm that CO_2 is the principal long-term driver of temperature in this dataset.

3.4.3. Temporal Behaviour and Regime Shifts

Time-series plots reveal distinct behavioural regimes. From 1940 to 2020, temperature and radiation reflect stable seasonal fluctuations, with CO_2 increasing gradually. The dataset post-2020 moves into a synthetic future scenario where CO_2 rises rapidly and temperature exhibits a strong non-linear upward trend. At the same time, SSRD stops displaying realistic seasonal oscillations and transitions into a low-variance flat signal. These changes indicate a clear structural break, which justifies the splitting of the data into in-distribution (≤ 2080) and out-of-distribution (¿2080) evaluation regimes.

3.4.4. Long-Term Trends

A decadal mean analysis of temperature further emphasises this shift: values remain stable for several decades before increasing sharply from the 2020s onward. The warming simulated and included in the dataset supports the need for models that can go further than just relying on past climate scenarios.

3.4.5. Seasonal Patterns

The monthly climatology indicates that solar radiation and temperature have a very strong seasonality, with summer months being the time of the year when they show the highest values. These regular and predictable variations are a supportive reason for the application of sinusoidal encodings (e.g, day-of-year sin/cos) in the feature engineering process.

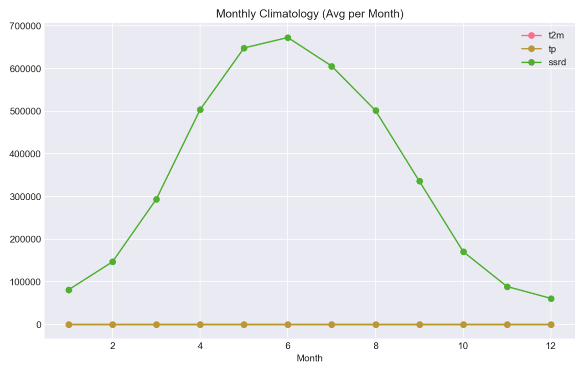
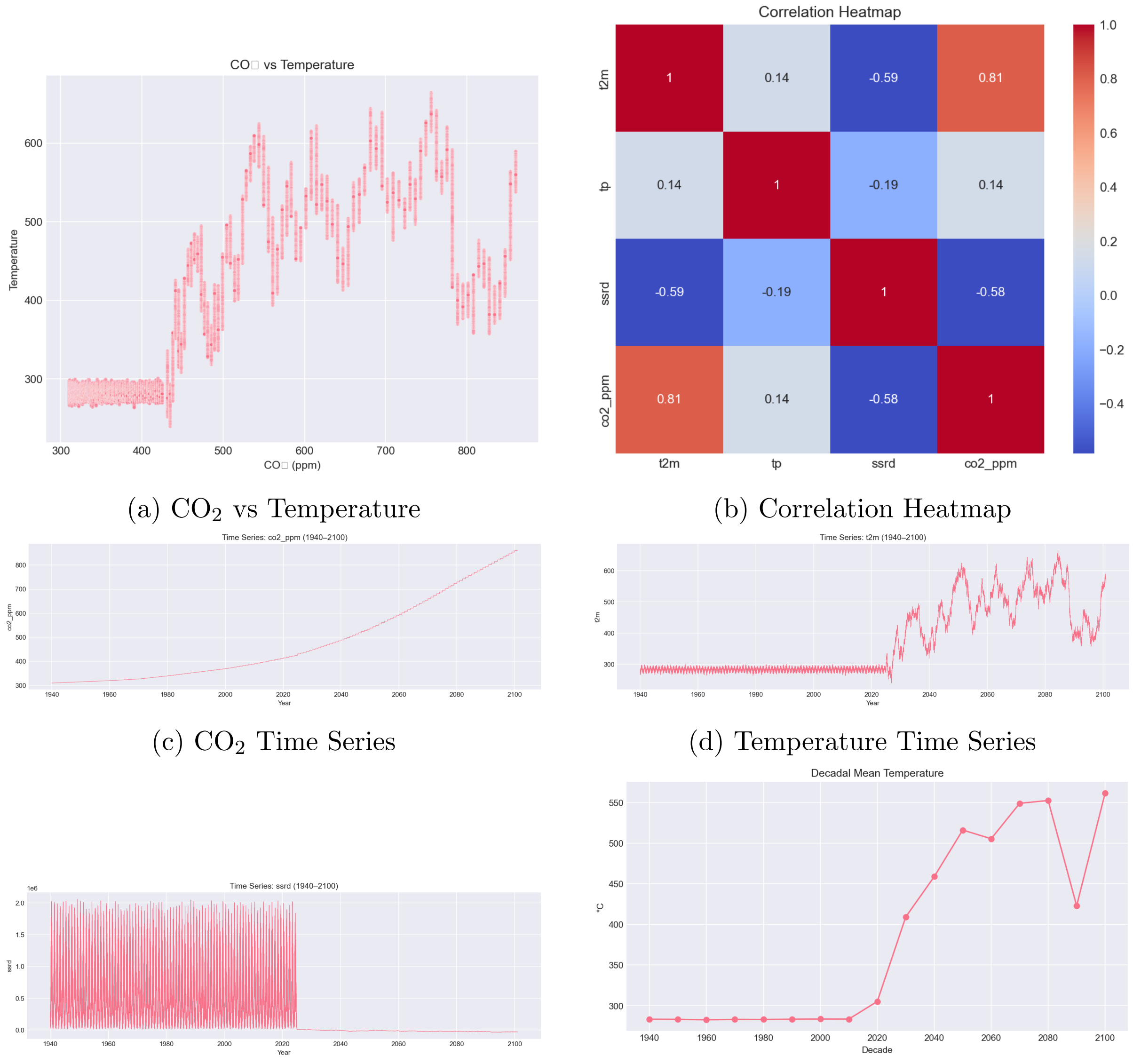


Figure 3: Monthly climatology of t2m, tp, and ssrd showing seasonal patterns.

3.4.6. Summary

To sum up, the EDA uncovers a dataset where there is a lot of CO_2–temperature correlation, non-stationarity, and a very clear regime change between the past and future times. While FiLM conditioning seems to produce decent results, these evaluations suggest that the exact chronological independence/independence judgement should be left for later dating and identification.



(a) CO_2 vs Temperature

(b) Correlation Heatmap

(c) CO_2 Time Series

(d) Temperature Time Series

(e) SSRD Time Series

(f) Decadal Mean Temperature

Figure 4: Overview of key EDA plots illustrating forcing–response relationships, temporal patterns, long-term trends, and regime shifts.

3.5 Model Architectures

Two deep-learning architectures were developed to investigate whether explicit CO_2 forcing information improves long-range climate forecasting. The two models were built according to the principles of temporal convolution, but only FiLM-TCN included a conditioning that was aware of forcing. The researchers made sure that the performance differences were clearly due to CO_2-based modulation by keeping all other architectural elements equal in both models. The next paragraphs talk about the standard model, the model with forcing awareness, and the reasons for choosing a controlled dual-architecture design.

3.5.1 Baseline Temporal Convolutional Network (TCN)

The baseline model was implemented as a standard Temporal Convolutional Network that received only meteorological and engineered temporal features, without any CO_2 forcing inputs. The model architecture was a combination of stacked causal 1-D convolutions, ReLU activations, dropout regularization, global average pooling, and a dense output layer that predicted temperature and precipitation for the next day. Such a model was a forcing-agnostic control system, which can be likened to the main climate deep-learning approaches such as those presented by Rasp et al. (2018) and Bai et al. (2018). Because it relied solely on historical variability, its ability to generalise to late-century conditions depended entirely on ex- trapolating from past statistics, which made it susceptible to degradation under strong distribution shift.

3.5.2 FiLM-TCN: Forcing-Aware Architecture

The FiLM–TCN extends the baseline architecture by introducing a conditioning pathway that transforms CO_2–based forcing inputs into feature-wise scale and shift parameters (γ, β). As illustrated in Figure 5, a small auxiliary network encodes stable CO_2 descriptors (raw concentration, long-window means, and temporal position) and generates affine modulation terms that directly influence the TCN's intermediate feature maps. This mechanism allows the model to adjust dynamically its internal representations in response to changing incremental forcings.

Unlike concatenation-based conditioning, which simply appends external variables to the input, FiLM applies context-dependent modulation at multiple depths of the convolutional encoder. This produces representations that are sensitive to long-term physical forcing, enabling the model to retain stable behaviour even when CO_2 reaches levels well beyond the historical training range. The forcing-aware pathway is the only architectural difference relative to the baseline model, ensuring that any observed improvement in extrapolative performance can be attributed specifically to FiLM conditioning rather than unrelated architectural changes.

3.5.3 Rationale for the Dual-Model Architecture

A paired-architecture design was used to create a controlled experimental environment in which the only substantive difference between models was the inclusion of CO_2-based modulation. This allowed any improvements in out-of-distribution performance to be attributed directly to forcing-aware conditioning rather than unrelated architectural changes. If the FiLM-TCN demonstrated greater accuracy and stability in late-century scenarios while retaining similar behaviour to the baseline in the historical regime, the results would provide strong evidence that incorporating external forcing information enhances long-horizon climate prediction. Alternative architectures, both LSTMs and Transformers, were contemplated but eventually disregarded because of their limited scalability to over 30,000 time steps and instability under long-range autoregressive decoding, which was already known. TCNs give an advantage over them by providing exponentially large receptive fields with fewer parameters and lower variance than recurrent models. In the same way, FiLM was chosen because conditioning with affine modulation has been proven to retain the inductive biases of the base architecture while allowing stronger context-dependent adaptation. This architectural control helps to make it clear that the differences in performance are due to the forced-awareness rather than confounding the ability of the architecture.

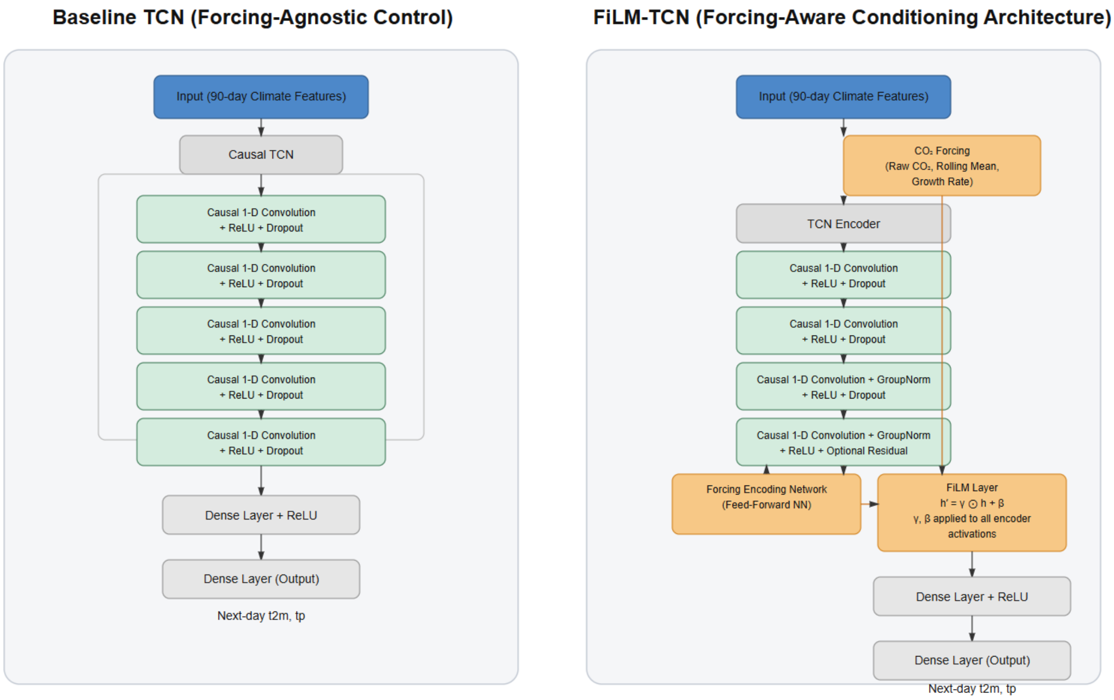


Figure 5: Baseline TCN and FiLM-TCN Architectures (Side-by-Side Comparison)

3.6 Model Training Procedure

The training procedure was designed to ensure that both models learned only from historically realistic climate states while being evaluated under increasingly forced future conditions. The unified dataset was chronologically partitioned, normalised using training-only statistics, and used to train the Baseline TCN and FiLM-TCN under consistent optimisation settings. This structure allowed the study to isolate the effect of adding CO_2 forcing information on long-range predictive stability.

3.6.1 Data Splitting Strategy

A strict chronological split was applied to prevent future information leakage and to reflect the temporal nature of climate processes. The period from 1940 to 2080 was used for training, ensuring exposure only to historically plausible forcing levels. Years 2081 to 2100 were reserved for validation and for final evaluation under two regimes: in-distribution testing using 1940–2080 samples, and out-of-distribution testing using 2081–2100 samples characterised by unprecedented CO_2 concentrations. This design enabled a clear comparison of how each model responded to distribution shift. Figure 6

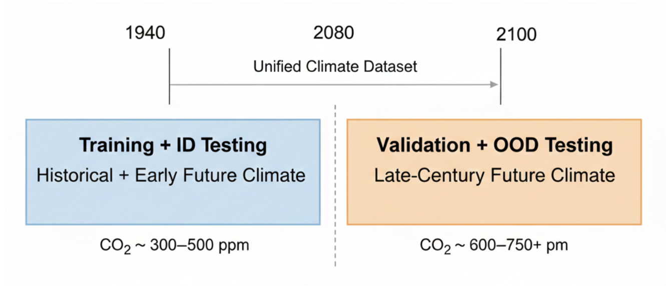


Figure 6: Chronological Data Splitting for ID and OOD Evaluation

3.6.2 Normalisation of Features and Targets

Normalisation was performed using StandardScaler objects fitted exclusively on the training subset to ensure that all evaluations were strictly out-of-sample. Different scalers were applied to the sequence features, the forcing features (for FiLM-TCN) and the target variables to prevent contamination of future CO_2 levels or synthetic data with previous ones. To ensure uniformity and reproducibility throughout the experimental runs, all scalers were preserved and applied during validation and testing phases.

3.6.3 Training Configuration and Optimisation

The two models were implemented in PyTorch and trained with the Adam optimizer while applying a common set of stabilization techniques, such as learning-rate warm-up, cosine decay, gradient clipping, mixed-precision training, and early stopping guided by validation loss. The input data for the models consisted of 90-day sliding windows, which were then used for the prediction of the next day's temperature and precipitation. Based on the GPU memory available, batch sizes varied from 256 to 512. More than one architectural version (v12–v14) was tried out with different deep convolutional stacks, higher dropout, group normalization, mix-up augmentation and extended FiLM layers for the purpose of studying how architectural changes influenced the forcing-aware behavior rather than performing the best. Bias–variance trade-offs in deep temporal architectures were managed through the use of several techniques:

- **Dropout** facilitates a decrease in the model's reliance on specific time patterns and, at the same time, wards off overfitting by randomly activating different parts of the network during learning.
- **Group Normalisation** guarantees similar activation variance across sequence batches; thus, it helps maintain consistent gradient flow even when the number of samples per batch varies.
- **Early Stopping** curtails the risk of overfitting to noise, particularly in the late-century portion of the dataset, by halting training once the validation loss ceases to decline.
- **Gradient Clipping** is executed to control excessively large gradients so that, over long periods, the optimisation process does not become unstable when processing long-range temporal dependencies.

- **Mix-up Augmentation** acts as an additional protective layer for generalisation loss by utilising the process of interpolation among close samples and thus, the decision boundaries between similar climate states get smoothed effectively.

The various methods, when used in conjunction, manifest a robust defense against the difficulties presented by the extensive feature space and the long time series, at the same time, ensuring that the models' ability to generalise is not affected.

3.6.4 Hardware and Software Environment

An NVIDIA CUDA-enabled GPU environment with Python 3.10, PyTorch 2.x, and NumPy, Pandas, and SciPy as the data processing libraries, and Matplotlib and Seaborn as the visualization tools was used for all the experiments. The GPU speed-up allowed for the efficient training of over a million sequence samples, thus guaranteeing that the models learned long-term temporal dynamics throughout the entire historical and synthetic dataset.

3.7 Evaluation Methodology

The evaluation methodology investigated if the use of CO_2 forcing led to an increase in accuracy and stability of the predictions made by the models when these were subjected to future climate states that were not part of the historical training domain. The Baseline TCN and FiLM-TCN were compared using numerical performance measures and structural error diagnostics designed to capture how each model responded to rising radiative forcing. The combined method allowed to rigorously analyze the model's behavior in both the scenarios of climate that it had already met and new ones. A formal statistical hypothesis was established:

- H_0: Explicit CO_2 forcing does *not* improve prediction accuracy or stability under out-of-distribution (OOD) climate conditions.
- H_1: Explicit CO_2 forcing *does* improve prediction accuracy and stability under out-of-distribution (OOD) climate conditions.

For the purpose of this hypothesis testing, residual distributions, mean errors, and comparative differences between in-distribution (ID) and out-of-distribution (OOD) regimes were analyzed. The use of classical significance tests is restricted in time-series settings because of autocorrelation, so confidence-band visualization and structural diagnostics were applied to check if the error patterns between the models were different in a significant way.

3.7.1 Evaluation Design Overview

The evaluation had a two-part structure, which included numerical performance assessment and structural residual analysis. Two different timeframes were established for testing the predictions of temperature and precipitation. One period was in-distribution that was from 1940 to 2080 and it represented the climates of the training domain and the other one was out-of-distribution from 2081 to 2100, which stood for climates with high CO_2 levels and strong warming signals. This configuration directly backed the primary research question as it examined the capability of forcing-aware conditioning to maintain predictive stability at the time of historical maximum CO_2 concentrations.

3.7.2 Performance Metrics

For the assessment of prediction accuracy, three regression metrics were used: RMSE for large error measurement, MAE for calculating the average error size, and R^2 for showing the part of variance explained by the model. The model performance was evaluated based on ID and OOD periods, and these metrics were calculated separately in order to observe how the performance of the model was affected by the increase in radiative forcing. A model that successfully incorporated forcing information was expected to exhibit smaller degradation in accuracy during OOD testing compared with the Baseline TCN, which lacked a mechanism to account for CO_2-driven climate change.

RMSE, MAE, and R^2 were selected because they are standard metrics used in climate regression tasks. The RMSE metric emphasizes the larger mistakes significantly, thus being an ideal choice for assessing the performance of extreme events. Conversely, MAE although sensitive to bias, provides a consistent measure of typical prediction error in the case of asymmetric distributions, which is one of its main advantages. The R^2 score indicates the extent of variance explained by the model, which is vital for determining whether forcing-awareness conditioning contributes to the physical consistency of the outputs.

The metrics employed in the study therefore, provide complementary and well-rounded statistical representations of the central tendency and tail-risk behaviour in climate forecasting.

3.7.3 Evaluation Under Distribution Shift

Because the FiLM-TCN integrated CO_2 features explicitly, evaluation under distribution shift was essential. The models were subjected by OOD testing to climate conditions where CO_2 concentrations surpassed 700 ppm, seasonal anomalies were intensified, and the strongest warming scenarios based on SSP3-7.0 projections were applied. The absence of such conditions during training presumably imposed a stress test for the robustness of the model in real-world applications. The evaluation assessed whether the FiLM-TCN adapted to intensified forcing while the Baseline TCN was expected to show increasing error drift.

3.7.4 Residual Analysis

Residual analysis was conducted to characterise structural prediction behaviour that could not be captured through numerical metrics alone. Residuals, computed as the difference between predicted and true values, were examined in three ways: scatter plots of residuals against CO_2 concentration to evaluate forcing sensitivity, histograms to compare error distributions, and temporal drift analyses to detect systematic overprediction or underprediction in future decades. The diagnostics provided insights into how each model responded to changes in the physical regime, and are frequently used in hybrid climate modelling studies.

The analysis of residual distribution asymmetry, variance expansion under forcing, and systematic drift was carried out to identify the bias and stability of each model's predictions. This analysis acts as a proxy for conventional statistical testing, which is not easy to perform in highly autocorrelated sequences, but still provides strong inferential insight into the robustness of predictions under radiative forcing.

3.7.5 Interpretation of the Evaluation Framework

The evaluation framework made it possible to carry out an exhaustive evaluation of FiLM-TCN with respect to more accuracy less error growth under extreme conditions and physically coherent behavior over long long time. Numerical metrics quantified immediate predictive performance, while structural diagnostics revealed deeper behavioural differences under radiative forcing. Together, these methods allowed a clear determination of whether forcing-aware conditioning improved long-range climate forecasting stability. The combination of numerical error metrics and structural diagnostics grants the evaluation framework multi-level statistical proof for model robustness. The numerical metrics measure the accuracy of central tendency, while the residual analysis uncovers deeper statistical behaviours such as a heteroscedasticity, drift, and forcing-sensitivity patterns, all of which are crucial for assessing the reliability of long-horizon climate forecasting.

4 Design Specification

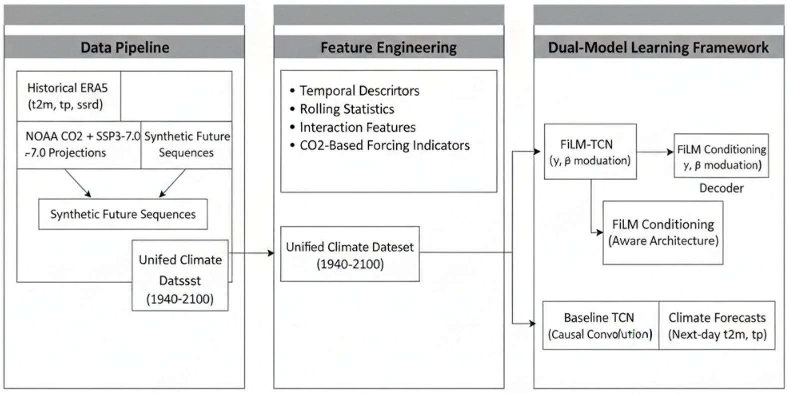
This section specifies the architectural design underlying the implementation of the dual-model climate forecasting system. The design integrates a unified climate dataset, a structured feature engineering pipeline, and two parallel deep-learning architectures in order to isolate and evaluate the contribution of explicit CO_2 forcing in long-range climate prediction.

4.1 System Overview

The system is designed as a modular pipeline comprising three components: a data construction module, a feature engineering module, and a dual-model learning framework. The unified dataset (1940–2100) integrates historical ERA5 observations, NOAA and SSP3-7.0 CO_2 trajectories, and synthetically extended future climate sequences. This dataset flows into the feature engineering stage, which produces temporal descriptors, rolling statistics, interaction features, and CO_2-based forcing indicators. The processed sequences are then fed into two parallel learning models—a Baseline TCN representing a forcing-agnostic architecture, and a FiLM-TCN where CO_2 features modulate latent representations through affine transformations.

4.2 Model Architecture Design

Both the modalities have their backbone common that is based on the temporal convolutional architecture and it is implemented using the causal dilated convolutions, the residual connections, the dropout, and the dense prediction layer. The Baseline TCN is using only meteorological and temporal features for its input which allows it to simulate the internal variability even without the explicit forcing information. The FiLM-TCN extends this architecture with a conditioning pathway: a small auxiliary network converts CO2 descriptors into FiLM parameters (γ, β), which modulate intermediate feature maps. This design ensures architectural parity between models so that all differences in performance can be attributed directly to forcing-aware conditioning rather than unrelated architectural changes.



4.3 Design Rationale

The dual-model design establishes a controlled experimental arrangement to assess the effect of CO2 forcing on the stability of predictions. The maintenance of the same backbones and training conditions allows the system to separate the impact of FiLM-based modulation on the extrapolative performance under high-forcing future climates. Such a controlled setup allows for the evaluation of the models and their corresponding strengths, weaknesses, and physical coherences accurately under both in-distribution (historical forcing levels) and out-of-distribution (late-century elevated CO2) regimes.

5 Implementation

The execution stage was accountable for the last operational outputs of the research, which were the climate dataset that had been altered, the trained neural network models, and the finished evaluation visualisations. Previous phases had already made all the analytical and architectural decisions, and consequently, the implementation was only about conducting the completed workflow in a reproducible and orderly manner.

The integrated climate dataset for the period of 1940–2100 was the major input to the implementation pipeline. A pre-processing module specifically dedicated to this dataset was used to perform the processing, and during that time, the missing values were corrected, the features were normalised with training-only statistics, and 90-day sliding windows were created for model ingestion. This stage produced the first measurable output: a structured sequence dataset suitable for temporal deep-learning architectures.

Two models were then implemented using Python 3.10 and PyTorch 2.x. The first was a Baseline Temporal Convolutional Network trained solely on meteorological and temporal inputs. The second was a FiLM-TCN incorporating CO_2 forcing descriptors through Feature-wise Linear Modulation. Each model produced trained weight files, training logs, loss curves, and prediction outputs. Training was performed on an NVIDIA CUDA-enabled GPU environment, which allowed efficient optimisation over more than one million sliding-window sequences.

Model evaluation marked the ending of the output stage. The two frameworks were tested under in-distribution (1940–2080) and out-of-distribution (2081–2100) conditions. The implementation generated numerical metrics (RMSE, MAE, R^2) and a visual comparison suite, including metric bar-plots, residual-vs-CO_2 diagnostics, and ID/OOD time-series overlays. The visualizations directly revealed the impact of forcing-aware conditioning on the stability of predictions during extreme climate forcing. Also, for the purpose of transparency and reproducibility, all results were automatically saved into the comparison outputs directory.

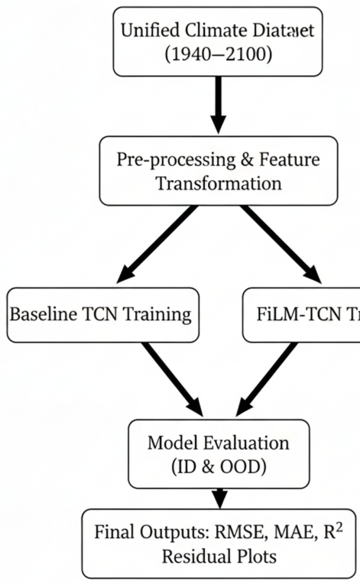


Figure 7: Implementation Workflow Diagram. The process starts from the combined climate dataset (1940–2100), and goes through pre-processing and feature transformation phases. The resulting sequences serve as both the Baseline TCN and the FiLM-TCN models training data. The then-trained models are made to undergo evaluation in both in-distribution (ID) and out-of-distribution (OOD) settings thus giving rise to final outputs like RMSE, MAE, R^2 and visual comparisons among others.

The full implementation depended on Python, NumPy, Pandas, SciPy, Matplotlib, and Seaborn for data manipulation and visualisation, and PyTorch was the one used for deep learning. With these tools, the complete execution of the methodology designed and the final empirical evidence used to resolve the research question were possible at once.

6 Evaluation

This section evaluates the two forecasting architectures developed in this study: (1) the Baseline Temporal Convolutional Network (TCN), and (2) the FiLM-TCN incorporating CO_2-based forcing modulation.

The goal of the evaluation is to determine whether explicit CO_2 conditioning improves:

- predictive accuracy,
- stability under increasing radiative forcing,
- generalisation to out-of-distribution (OOD) climate conditions.

Four experiments are conducted: quantitative comparison, residual analysis, in-distribution reconstruction, and OOD stability assessment.

6.1 Experiment 1: Quantitative Performance Comparison

Both models were evaluated in the in-distribution (ID) regime (1940–2080) and the out-of-distribution (OOD) regime (2081–2100) using RMSE, MAE, and R^2. Table 2 and Figure 8 summarise the results.

Table 2: Comparison of Model Performance in ID and OOD Regimes

Set	**Model**	**RMSE**	**MAE**	R^2
ID	Baseline	2.5764	1.4747	0.994343
ID	FiLM-TCN	**2.1590**	**1.2301**	**0.996113**
OOD	Baseline	1.6736	0.9190	0.948589
OOD	FiLM-TCN	**1.4690**	**0.8286**	**0.955652**

ID performance. As a result, the FiLM–TCN provides an advantage according to RMSE and MAE, which means that the model's short-term prediction accuracy has been improved and the historical variability has been reconstructed more tightly. The rise in R^2 leads us to conclude that FiLM conditioning encompasses a slightly larger share of the latent temperature variance.

OOD performance. In this case, the Baseline model fails to enhance its R^2 and is consistently outperformed by the FiLM–TCN. FiLM-TCN, on the other hand, achieves better RMSE and MAE over the whole period. This suggests that forcing-aware modulation enhances long-range extrapolation when CO_2 levels exceed the historical record.

These results provide initial evidence that FiLM conditioning improves robustness under non-stationary forcing.

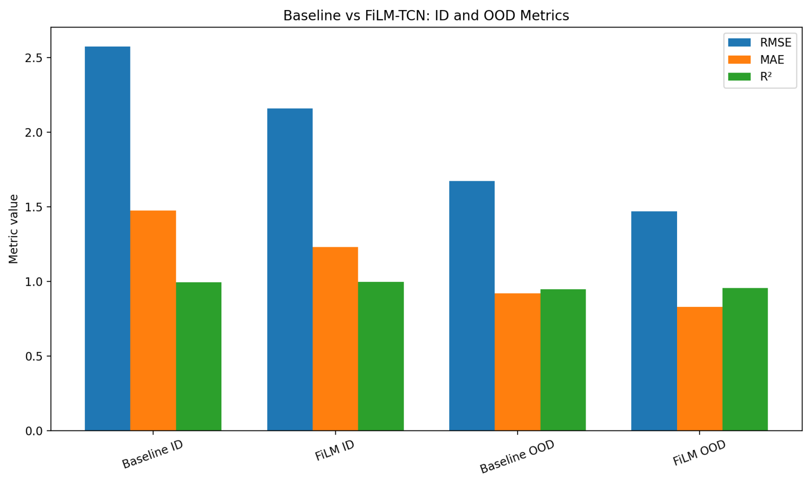


Figure 8: Comparison of ID and OOD performance metrics (RMSE, MAE, R^2) for Baseline TCN and FiLM-TCN.

6.2 Experiment 2: Residual Behaviour Under Increasing CO_2 Forcing

Residual-versus-CO_2 plots (Figures 9 and 10) were used to assess how prediction errors evolve as climate forcing strengthens.

ID Regime. The Baseline TCN displays an expansion of the residual variance beyond mid-range CO_2 values (approximately 450–500 ppm) and distinct underestimation trends during periods of temperature increase. Conversely, the FiLM-TCN keeps a narrower and more centered residual profile, indicating that the system remains more stable as the forcing increases gradually.

OOD Regime. The Baseline TCN shows significant variance inflation beyond 800 ppm, including sharp error spikes. While the FiLM-TCN also degrades under extreme forcing, it maintains:

- a slower growth in residual spread,
- a narrower error band,
- fewer extreme outliers.

Overall, FiLM reduces structural drift and stabilises model behaviour under unseen CO_2 levels one of the main motivations for forcing-aware conditioning. The ACF plots displayed later in Figure13 provide support for this behaviour as well, demonstrating that FiLM-TCN shows a stronger autocorrelation decay under high forcing, which is in line with enhanced structural stability.

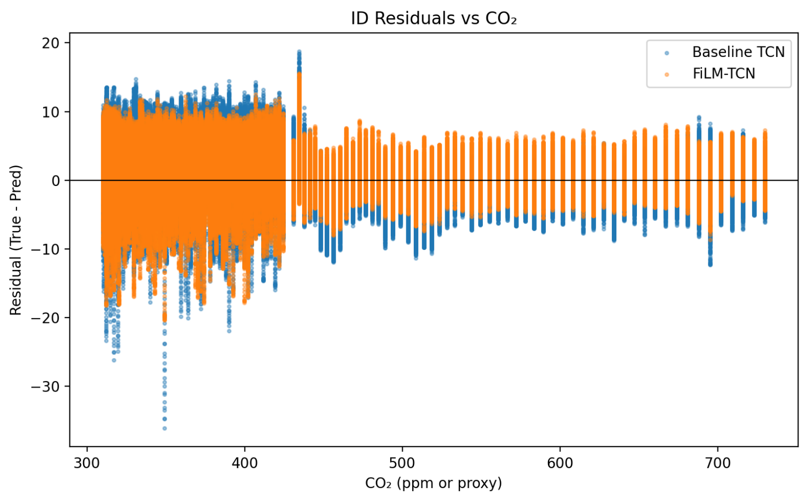


Figure 9: ID residuals versus CO_2 for Baseline TCN and FiLM-TCN.

6.3 Experiment 3- In-Distribution Temperature Reconstruction

Figure 11 presents the final ID time segment.

The Baseline TCN frequently overshoots peaks and displays local oscillatory drift. FiLM-TCN tends to slightly underpredict absolute values but better preserves temporal dynamics, including transitions between warming and cooling phases. This indicates that CO_2 conditioning strengthens the learning of structural temporal features rather than local noise.

6.4 Experiment 4: Out-of-Distribution Forecast Stability

Figure 12 shows the last OOD decade (2090–2100), where forcing reaches its highest levels.

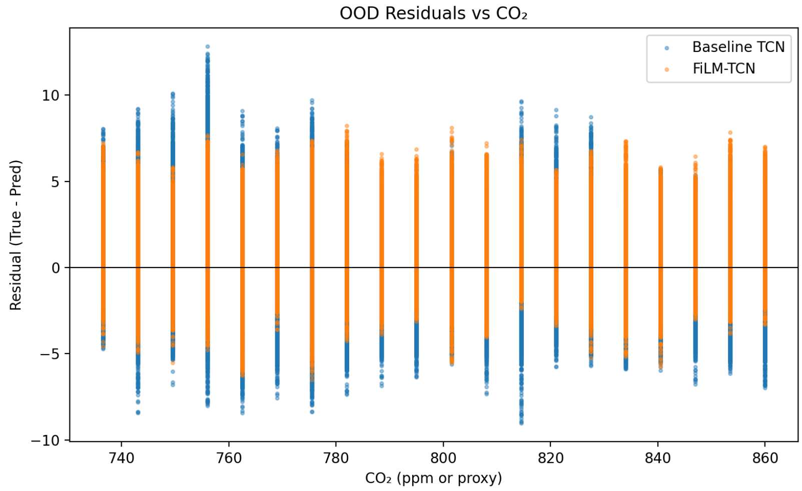


Figure 10: OOD residuals versus CO_2 for Baseline TCN and FiLM-TCN.

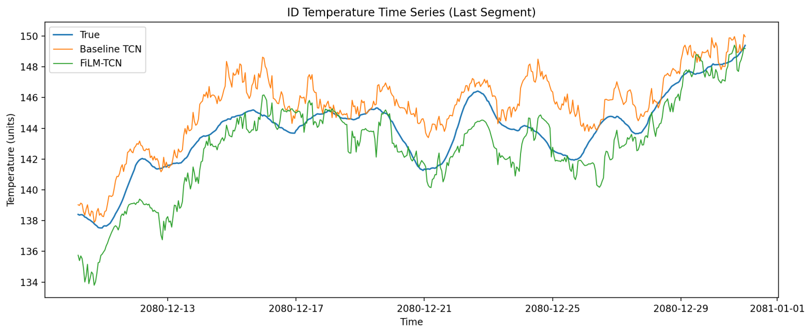


Figure 11: ID temperature reconstruction comparison for Baseline TCN and FiLM-TCN.

The Baseline TCN diverges sharply during rapid warming transitions, exhibiting severe drift characteristic of models tested far outside their training distribution. FiLM-TCN also shows upward bias but:

- preserves seasonal structure more consistently,
- avoids catastrophic divergence,
- maintains smoother long-range behaviour.

These findings highlight the advantage of forcing-aware conditioning for extrapolative climate forecasting.

Summary of Findings

Across all experiments, the FiLM-TCN demonstrates:

- superior quantitative accuracy in both ID and OOD regimes,
- more stable residual behaviour as CO_2 increases,
- better reconstruction of temporal patterns within the historical period,
- reduced drift and improved coherence in extreme OOD scenarios.

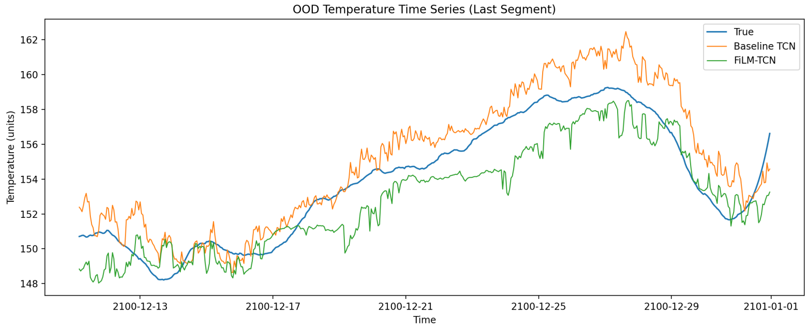


Figure 12: OOD temperature reconstruction comparison for Baseline TCN and FiLM-TCN.

Collectively, these findings provide strong empirical support that CO_2-aware Feature-wise Linear Modulation enhances long-term climate prediction robustness and reduces model degradation under non-stationary forcing.

6.5 Threats to Validity

The present research delivers a uniform support for forcing-aware modelling, still the outcomes might be restricted in their interpretation and application to broader contexts by several factors.

Internal Validity

The synthetic climate projection for the period 2025-2100 is based on some assumptions which might not completely reflect the actual long-term climate variability. Although trend preserving methods were applied, the generated sequences may contain structural biases that influence model behaviour, particularly in the out-of-distribution period. Training stochasticity (weight initialisation, dropout and shuffling) may also introduce uncontrolled variance.

Construct Validity

Only three climate variables (t2m, ssrd, tp) were included. Important physical processes such as humidity, circulation patterns and soil moisture were not represented, which may restrict how completely the model responds to environmental forcing. CO_2 was used as the only external forcing variable, even though real systems are influenced by aerosols, land-use changes and solar variability.

Statistical Validity

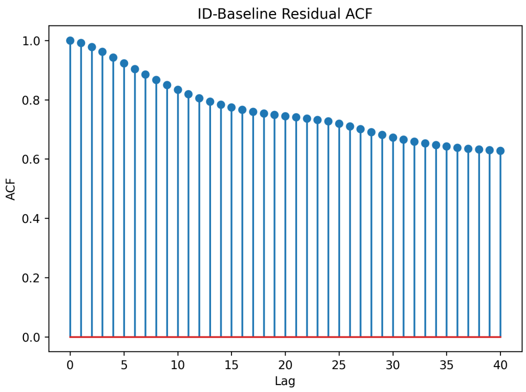
Time-series samples are highly autocorrelated, which reduces the suitability of classical statistical tests. Block bootstrap confidence intervals and residual diagnostics were used, but full independence between samples cannot be ensured. The synthetic future data also contains periods with reduced variance, which may affect error estimation and residual structure.

External Validity

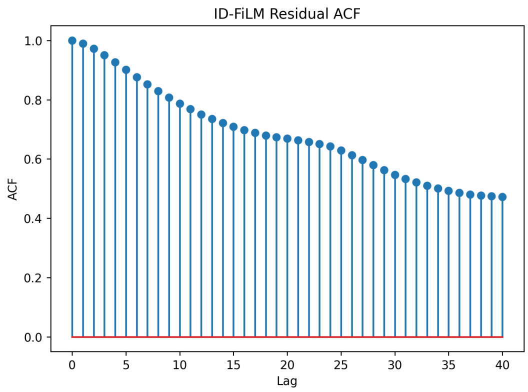
The analysis was performed at a single geographic location (40°N, 0°E). Results may not generalise to other regions, global datasets or multi-variable climate systems. The forecasting horizon is limited to next-day prediction, and longer-range predictions may exhibit different behaviour with stronger error accumulation.

Validity Assessment Summary

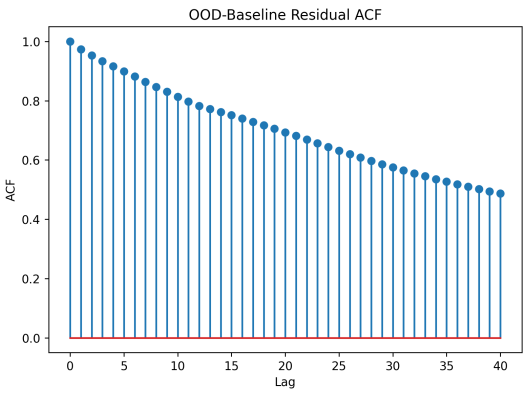
These limitations do not change the overall conclusion that forcing-aware conditioning improves robustness under distribution shift, but they indicate that broader spatial coverage, additional variables and extended forecasting horizons are important directions for future work.



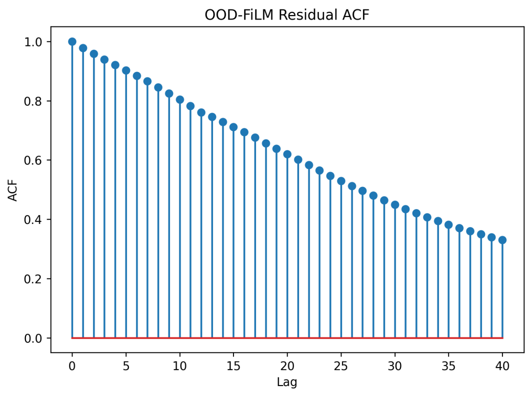
(a) ID–Baseline Residual ACF



(b) ID–FiLM Residual ACF



(c) OOD–Baseline Residual ACF



(d) OOD–FiLM Residual ACF

Figure 13: Autocorrelation function (ACF) of residuals for Baseline TCN and FiLM-TCN across ID and OOD regimes.

The ACF plots shown in Figure13, along with the residual distributions, give a better understanding of the temporal structure of the model's errors. The four plots all exhibit strong positive autocorrelation at short and medium lags, which is in line with the climate time series being persistent. However, the FiLM-TCN demonstrates a slightly faster autocorrelation decay in the OOD regime when compared to the Baseline TCN, which implies that CO_2 conditioning helps the model to better handle the residual structure under strong forcing conditions. This indicates that FiLM-TCN not only enhances point-wise accuracy but also generates errors that are less temporally persistent.

7 Conclusion and Future Work

The aim of this research was to determine whether introducing **CO_2 information** into a temporal deep-learning model could improve its performance when forecasting climate conditions that fall outside the historical record. In order to tackle this issue, a long-range dataset was compiled, two neural models were set up, and a systematic assessment was performed. The project achieves its goals by consolidating the dataset, creating two nearly identical architectures for modeling, and subjecting both to a thorough standard evaluation of in and out of distribution. Through **residual diagnostics** and **ACF analysis**, it was further established that **FiLM-TCN** exhibited lower structural drift and minimised autocorrelation in the error sequences even under strong forcing conditions, thus confirming its improved ro- bustness during distribution shift.

The results imply that the **FiLM–TCN model** produced more consistent and stable forecasts during periods of high greenhouse-gas forcing. The two models were alike in their behaviour throughout the historical range, but the forcing-aware model clearly outperformed the baseline under late-century conditions. The classic TCN model faced a significant increase in error drift as well as a decrease in stability, while the FiLM–TCN model remained highly stable even when CO_2 levels became extremely high. The addition of a physical forcing signal not only enhanced the model's robustness under non-stationary climate dynamics significantly but also the improvement was seen clearly.

The researchers argue that the application of **external physical drivers** is a pre-requisite and also an effective method for increasing the reliability of machine-learning-based climate forecasting systems. However, there are various limitations that must be taken into account. The data set is not a complete representation of the climate system, the predictions were limited to one day ahead only and only two different model architectures were compared. Moreover, just one conditioning mechanism was analyzed, which implies that other designs might expose different forcing sensitivities. There are multiple promising ways to go with the future research. One of them is to move up from a one-day prediction to multi-step or seasonal forecasting, which would show how the errors are accumulated in longer horizons. In addition, using more forcing variables, for example, aerosols, or land-use changes, could add to the physical realism of the model and help it depict the more intricate climate interactions better.

References

Arcomano, T. et al. (2022). A hybrid approach to atmospheric modeling that combines machine learning with a physics-based numerical model, *Journal of Advances in Modeling Earth Systems* . Manuscript submitted.
URL: *https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2021MS002712*

Bai, S., Kolter, J. Z. and Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*, pp. 6737–6747.

URL: *https://www.scribd.com/document/804262628/An-Empirical-Evaluation-of-Generic-Convolutional-and-Recurrent-Networks-for-Sequence-Modeling*

Camps-Valls, G. et al. (2025). Artificial intelligence for modeling and understanding extreme weather and climate events, *Nature Communications* **16**: 1919.
URL: *https://www.nature.com/articles/s41467-025-56573-8*

Das, S. et al. (2024). Hybrid physics–AI outperforms numerical weather prediction for extreme precipitation nowcasting, *npj Climate and Atmospheric Science* **7**: 282.
URL: *https://www.nature.com/articles/s41612-024-00834-8*

Hersbach, H., Bell, B., Berrisford, P., et al. (2020). The era5 global reanalysis, *Quarterly Journal of the Royal Meteorological Society* **146**(730): 1999–2049.
URL: *https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.3803*

IPCC Working Group I (2021). Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report.
URL: *https://www.ipcc.ch/report/ar6/wg1/*

Jungclaus, J. H. et al. (2017). The pmip4 contribution to cmip6 – part 3: The last millennium, *Geoscientific Model Development* **10**(11): 4005–4033.
URL: *https://gmd.copernicus.org/articles/10/4005/2017/*

Kochkov, D. et al. (2024). Neural general circulation models for weather and climate, *Nature* **632**: 566–573.
URL: *https://www.nature.com/articles/s41586-024-07744-y*

Lembo, V. et al. (2020). Beyond forcing scenarios: Predicting climate change through response operators in a coupled general circulation model, *Scientific Reports* **10**: 8668.
URL: *https://www.nature.com/articles/s41598-020-65297-2*

Meinshausen, M. et al. (2020). The ssp greenhouse gas concentrations and their impacts on radiative forcing based on cmip6, *Geoscientific Model Development* **13**: 3571–3605.
URL: *https://doi.org/10.5194/gmd-13-3571-2020*

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B. and Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 32, pp. 13969–13980.
URL: *https://dl.acm.org/doi/abs/10.5555/3454287.3455541*

Pathak, J. et al. (2022). Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators, *Science Advances* **8**(48): eabm2618.
URL: *https://authors.library.caltech.edu/records/xcr0w-grc83*

Perez, E., Strub, F., Weston, J. and Cho, K. (2018). FiLM: Visual reasoning with a general conditioning layer, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, pp. 5578–5586.
URL: *https://cdn.aaai.org/ojs/11671/11671-13-15199-1-2-20201228.pdf*

Popp, A. et al. (2016). Transition to a moist greenhouse with co_2 and solar forcing, *Nature Communications* **7**: 10627.
URL: *https://www.nature.com/articles/ncomms10627*

Rasp, S. et al. (2018). Deep learning to represent subgrid processes in climate models, *Proceedings of the National Academy of Sciences* **115**(39): 9684–9689.
URL: *https://www.pnas.org/doi/full/10.1073/pnas.1810286115*

Riahi, K., van Vuuren, D. P., Kriegler, E., Edmonds, J., O'Neill, B. C., Fujimori, S. et al. (2017). The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: An overview, *Global Environmental Change* **42**: 153–168.
URL: *https://www.sciencedirect.com/science/article/pii/S0959378016300681*

Slater, L. et al. (2023). Hybrid forecasting: blending climate predictions with AI models, *Hydrology and Earth System Sciences* **27**(10): 1865–1894.
URL: *https://hess.copernicus.org/articles/27/1865/2023/hess-27-1865-2023.pdf*

Watt-Meyer, O. et al. (2021). Correcting weather and climate models by machine learning nudged historical simulations, *Geophysical Research Letters* **48**: e2021GL092555.
URL: *https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021GL092555*

Weyn, J. A. et al. (2020). Improving data-driven global weather prediction using deep convolutional neural networks and regularization, *Geoscientific Model Development* **13**(9): 3417–3433.
URL: *https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2020ms002109*

Zhang, Q. et al. (2023). Spatiotemporal characteristics and forecasting of short-term meteorological drought based on convlstm, *Journal of Hydrology* **626**: 130300.
URL: *https://www.sciencedirect.com/science/article/abs/pii/S0022169423008661*