Director
Dr. Oriol Tintó Prims
Departament de Ciències de la Terra (BSC)

Tutor
Dr. Ibério Ribeiro Moreira
Departament de Ciència de Materials i Química Física (UB)

Màster oficial en

Modelització Computacional Atomística i Multiescala en Física, Química i Bioquímica

Treball Final de Màster

Quantificació de la incertesa en prediccions climàtiques d'alta resolució utilitzant mètodes d'IA

Cuantificación de la incertidumbre en predicciones climáticas de alta resolución mediante métodos de IA

AI-assisted uncertainty quantification in high-resolution climate predictions

Aina Gaya i Àvila

Juliol 2025

UNIVERSITAT DE BARCELONA
B:KC Barcelona Knowledge Campus
Campus d'Excel·lència Internacional
UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

I would like to express my sincere gratitude to Oriol Tintó Prims for supervising my Master's thesis. Throughout these months, we have continuously exchanged ideas, thoughts, and engaged in discussions that have kept me motivated. I am deeply grateful for his guidance and support.

I would also like to extend my appreciation to everyone in the Computational Earth Sciences group at BSC, especially Miguel Castrillo and Francesca Macchia, who placed their trust in me from the moment we met. I am equally grateful to all those with whom I had the opportunity to discuss this work, who generously shared their time and insights to help define and refine my research: Amirpasha Mozzafari, Alejandro Peraza, Lluís Palma, Eneko Martín, Llorenç Lledó, Paolo Davini, Jost von Hardenberg and Thomas Rackow. I owe a tremendous debt of gratitude to my colleagues, particularly Alba Martínez, Ariadna Batalla, Carmen Piñero, Celia Ferruz, Johanna Gehlen, Maitane Fariñas, Marta Alerany, Paula Serrano, and Lorena Gallardo, among many others, for their constant support and for making my time at the office enjoyable in addition to productive. I want to extend my gratitude to those who took time to read and revise this manuscript: Okke van Eck, Kat Grayson, Ivan Alsina, Francesc Roura and Iker Gonzalez.

I wish to acknowledge my friends, who have supported me despite being scattered across the globe: some near Barcelona, while others are in Utrecht, Copenhagen, Madrid, Jena, and Munich. A special mention must go to Francesc Sabater for his ability to bring lightness to even the most demanding days, constantly reminding me that humor is not just welcome but essential during challenging times.

Finally, I extend my deepest gratitude to my mother, father, and sister, without whom none of this would have been possible, and to my grandfathers, from whom I learned everything that truly matters in life.

Títol: **Quantificació de la incertesa en prediccions climàtiques d'alta resolució utilitzant mètodes d'IA**

Estudiant: Aina Gaya i Àvila

Data: Juliol 2025

Director: Dr. Oriol Tintó Prims
Departament de Ciències de la Terra (BSC)

Tutor: Dr. Ibério Ribeiro Moreira
Departament de Ciència de Materials i Química Física (UB)

L'adaptació i mitigació del canvi climàtic requereix informació climàtica global d'alta resolució. Les iniciatives actuals que generen aquesta informació (Destination Earth) manquen d'estimacions d'incertesa, ja que els enfocaments tradicionals per a la quantificació d'incertesa, basats en la predicció per conjunts (executar múltiples simulacions amb paràmetres o condicions inicials lleugerament diferents) són computacionalment costosos. Aquest estudi presenta una alternativa a aquest problema: un marc de treball per quantificar la incertesa utilitzant models basats en intel·ligència artificial (IA), que són més econòmics d'executar.

S'ha desenvolupat un marc de treball que integra dos models d'IA d'última generación, NeuralGCM i AIFS, i permet inicialitzar-los a partir de dades provinents de Destination Earth. Emprant aquest models menys costosos podem generar conjunts per tal de quantificar la incertesa de les dades. El marc és agnòstic al model i modular, facilitant la integració de diversos models de predicció basats en IA, així com l'addició d'altres fonts de condicions inicials. L'arquitectura utilitza contenidors per assegurar la portabilitat i reproducibilitat entre diferents entorns de computació d'alt rendiment, i s'integra amb Autosubmit, el gestor de fluxos de treball desenvolupat al Barcelona Supercomputing Center.

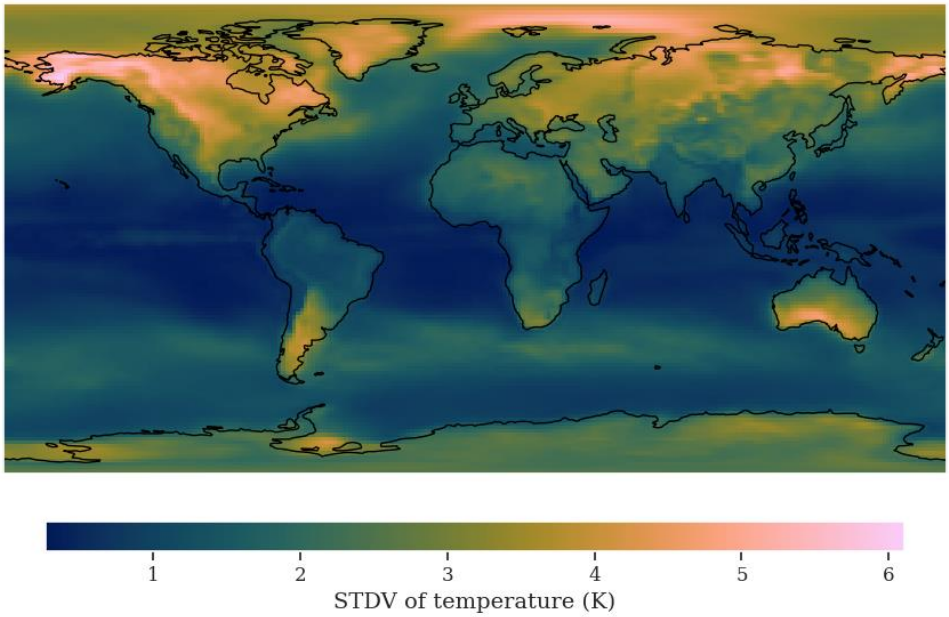

Figura 1: Desviació estàndard obtinguda amb NeuralGCM durant Gener de 2039. Les zones amb una desviació estàndard major són aquelles en les que hi ha més variabilitat, i per tant, major incertesa.

Aquest marc s'ha emprat per correr diverses simulacions de 10 dies a MareNostrum5. S'ha estudiat tant el comportament dels models, per comprobar si pateixen algún tipus de biaix en executar-los fora de les condicions sota les quals han estat entrenats, com la viabilitat d'emprar aquests models per quantificar la incertesa de les dades emprades com a condicions inicials. Els resultats experimentals demostren diferències significatives en el rendiment entre els models avaluats. Sota condicions climàtiques més càlides (escenari SSP3-7.0), AIFS mostra una deriva sistemàtica cap a un estat més fred, mostrant una tendència a retornar a les condicions de l'estat d'entrenament. En contrast, NeuralGCM manté la consistència física amb les simulacions de Destination Earth, produint trajectòries que segueixen les del model de referència.

L'anàlisi de la incertesa revela que NeuralGCM genera distribucions de conjunts consistents des de les quals es pot estimar la incertesa de les dades climàtiques d'alta resolució. La quantificació de la incertesa s'avalua mitjançant mètriques establertes: la dispersió (Spread), l'error quadràtic mitjà (RMSE) i la relació dispersió-error (SER). S'avalua la incertesa de les simulacions pel període de Gener de 2039, estimant quines son les regions on hi ha més incertesa durant aquest interval de temps.

En conclusió, aquesta recerca demostra la viabilitat de la quantificació d'incertesa assistida per IA per a projeccions climàtiques d'alta resolució, proporcionant un pont entre els mètodes tradicionals de conjunts computacionalment costosos i la necessitat d'estimacions d'incertesa pràctiques. El marc desenvolupat és una eina que pot ser emprada per a futures investigacions i aplicacions.

Título: **Cuantificación de la incertidumbre en predicciones climáticas de alta resolución mediante métodos de IA**

Estudiante: Aina Gaya Àvila

Fecha: Julio 2025

Director: Dr. Oriol Tintó Prims
Departamento de Ciencias de la Tierra (BSC)

Tutor: Dr. Ibério Ribeiro Moreira
Departamento de Ciencia de Materiales i Química Física (UB)

La adaptación y mitigación del cambio climático requiere información climática global de alta resolución. Las iniciativas actuales que generan esta información (Destination Earth) carecen de estimaciones de incertidumbre, ya que los enfoques tradicionales para la cuantificación de incertidumbre, basados en la predicción por conjuntos (ejecutar múltiples simulaciones con parámetros o condiciones iniciales ligeramente diferentes) son computacionalmente costosos. Este estudio presenta una alternativa a este problema: un marco de trabajo para cuantificar la incertidumbre utilizando modelos basados en inteligencia artificial (IA), que son más económicos de ejecutar.

Se ha desarrollado un marco de trabajo que integra dos modelos de IA de última generación, NeuralGCM y AIFS, y permite inicializarlos a partir de datos provenientes de Destination Earth. Empleando estos modelos menos costosos podemos generar conjuntos para cuantificar la incertidumbre de los datos. El marco es agnóstico al modelo y modular, facilitando la integración de diversos modelos de predicción basados en IA, así como la adición de otras fuentes de condiciones iniciales. La arquitectura utiliza contenedores para asegurar la portabilidad y reproducibilidad entre diferentes entornos de computación de alto rendimiento, y se integra con Autosubmit, el gestor de flujos de trabajo desarrollado en el Barcelona Supercomputing Center.

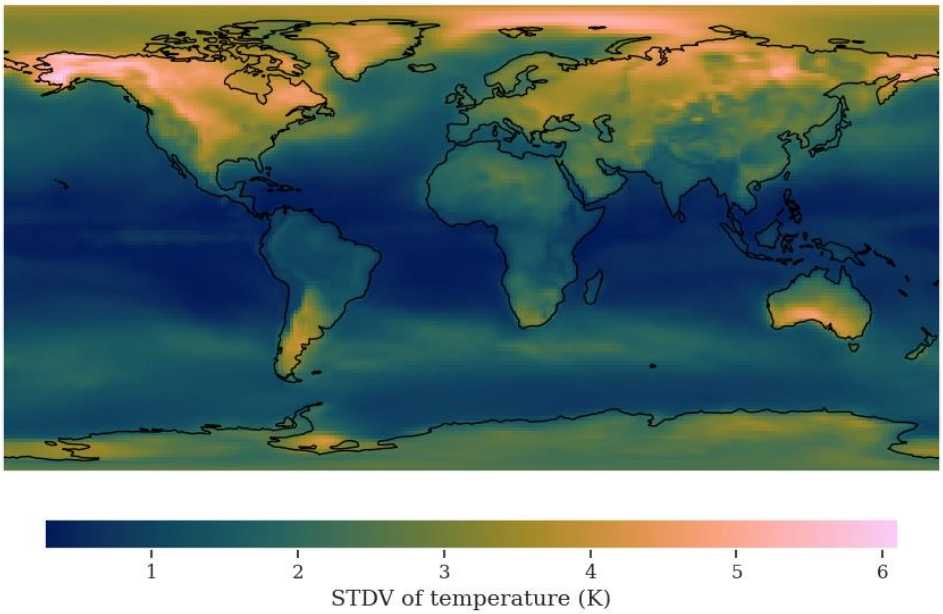

Figura 1: Desviación estándar obtenida con NeuralGCM durante Enero de 2039. Las zonas con mayor desviación estándar son aquellas en las que hay más variabilidad, y por tanto, mayor incertidumbre.

Este marco se ha empleado para correr varias simulaciones de 10 días en MareNostrum5. Se ha estudiado tanto el comportamiento de los modelos, para comprobar si sufren algún tipo de sesgo al ejecutarlos fuera de las condiciones bajo las cuales han sido entrenados, como la viabilidad de emplear estos modelos para cuantificar la incertidumbre de los datos empleados como condiciones iniciales. Los resultados experimentales demuestran diferencias significativas en el rendimiento entre los modelos evaluados. Bajo condiciones climáticas más cálidas (escenario SSP3-7.0), AIFS muestra una deriva sistemática hacia un estado más frío, mostrando una tendencia a retornar a las condiciones del estado de entrenamiento. En contraste, NeuralGCM mantiene la consistencia física con las simulaciones de Destination Earth, produciendo trayectorias que siguen las del modelo de referencia.

El análisis de la incertidumbre revela que NeuralGCM genera distribuciones de conjuntos consistentes desde las cuales se puede estimar la incertidumbre de los datos climáticos de alta resolución. La cuantificación de la incertidumbre se evalúa mediante métricas establecidas: la dispersión (Spread), el error cuadrático medio (RMSE) y la relación dispersión-error (SER). Se evalúa la incertidumbre de las simulaciones para el período de enero de 2039, estimando cuáles son las regiones donde hay más incertidumbre durante este intervalo de tiempo.

En conclusión, esta investigación demuestra la viabilidad de la cuantificación de incertidumbre asistida por IA para proyecciones climáticas de alta resolución, proporcionando un puente entre los métodos tradicionales de conjuntos computacionalmente costosos y la necesidad de estimaciones de incertidumbre prácticas. El marco desarrollado es una herramienta que puede ser empleada para futuras investigaciones y aplicaciones.

Title: **AI-assisted uncertainty quantification in high-resolution climate predictions**

Student: Aina Gaya Àvila

Date: July 2025

Supervisor: Dr. Oriol Tintó Prims
Earth Sciences Department (BSC)

Tutor: Dr. Ibério Ribeiro Moreira
Materials and Physical Chemistry Department (UB)

Climate change adaptation and mitigation requires high-resolution global climate information. Current initiatives generating this information (Destination Earth) lack uncertainty estimates, as traditional approaches for uncertainty quantification, based on ensemble forecasting (running multiple simulations with slightly different parameters or initial conditions) are computationally expensive. This study presents an alternative to this problem: a framework for quantifying uncertainty using artificial intelligence (AI)-based models, which are more economical to run.

A framework has been developed that integrates two state-of-the-art AI models, NeuralGCM and AIFS, and allows them to be initialized from data coming from Destination Earth. Using these less expensive models we can generate ensembles to quantify data uncertainty. The framework is model-agnostic and modular, facilitating the integration of various AI-based prediction models, as well as the addition of other sources of initial conditions. The architecture uses containers to ensure portability and reproducibility across different high-performance computing environments, and integrates with Autosubmit, the workflow manager developed at the Barcelona Supercomputing Center.

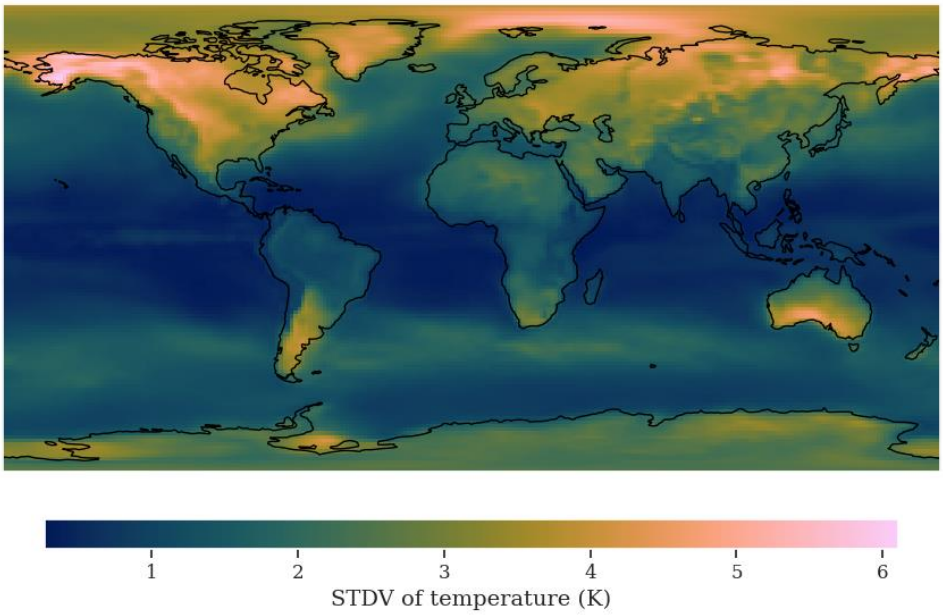

Figure 1: Standard deviation obtained with NeuralGCM during January 2039. The areas with a higher standard deviation are those in which there is more variability, and therefore, greater uncertainty.

This framework has been employed to run several 10-day simulations on MareNostrum5. Both the behavior of the models has been studied, to check if they suffer any type of bias when running them outside the conditions under which they have been trained, as well as the feasibility

of using these models to quantify the uncertainty of the data used as initial conditions. The experimental results demonstrate significant differences in performance between the evaluated models. Under warmer climatic conditions (SSP3-7.0 scenario), AIFS shows a systematic drift toward a colder state, showing a tendency to return to the training state conditions. In contrast, NeuralGCM maintains physical consistency with Destination Earth simulations, producing trajectories that follow those of the reference model.

The uncertainty analysis reveals that NeuralGCM generates consistent ensemble distributions from which the uncertainty of high-resolution climate data can be estimated. Uncertainty quantification is evaluated using established metrics: spread, root mean square error (RMSE), and spread-error ratio (SER). The uncertainty of the simulations is evaluated for the period of January 2039, estimating which regions have the most uncertainty during this time interval.

In conclusion, this research demonstrates the feasibility of AI-assisted uncertainty quantification for high-resolution climate projections, providing a bridge between traditional computationally expensive ensemble methods and the need for practical uncertainty estimates. The developed framework is a tool that can be used for future research and applications.

Descripció justificada de la contribució de l'estudiant al treball que es presenta.

Faig constar que la meva contribució en aquest treball ha consistit en:

- Anàlisi i valoració, conjuntament amb el director, de diversos models d'intel·ligència artificial que poguessin ser representatius. S'han emprat dos models pre-entrenats (NeuralGCM i AIFS) sense modificacions.

- Disseny i desenvolupament d'un workflow modular i escalable per a l'execució automatitzada de simulacions amb múltiples models basats en IA. Integració del workflow amb el gestor d'experiments Autosubmit, desenvolupat i utilitzat pel departament de Ciències de la Terra del BSC. Configuració detallada del sistema mitjançant fitxers YAML per tal de facilitar l'ús i reproduïbilitat del flux de treball. Automatització del procés d'execució, postprocessament i arxiu de dades amb scripts auxiliars integrats al workflow.

- Adaptació de contenidors existents per fer-los compatibles amb la infraestructura de MareNostrum5. Instal·lació de llibreries específiques.

- Preparació de condicions inicials a partir de dades de diverses fonts (ClimateDT i ERA5), incloent: Conversió i transformació de formats (Healpix → regular lat-lon, GRIB, Zarr, NetCDF) interpolació vertical entre diferents nivells de pressió per ajustar-se als requeriments de cada model, inferència de variables no disponibles (ex. ciwc), amb aproximacions físiques i factors d'escalat, estandardització de les dades d'entrada segons els requeriments específics de cada model, tenint en compte el conjunt de variables, resolucions i metadades.

- Execució de centenars de simulacions a la partició accelerada (GPU) de MareNostrum5, incloent: Simulacions de 10 dies i mesos sencers amb inicialitzacions diàries, generació de ensembles amb 10 membres per a model estocàstic. Generació d'un volum considerable de dades (2.6 TB) i gestió de l'espai de treball en entorns compartits HPC. Juntament amb el tutor, es va sol·licitar un projecte especific per tenir recursos dedicats a aquest projecte.

- Estimació del temps d'execució i consum energètic per simulació, mitjançant la llibreria EAR.

- Anàlisi de resultats amb enfoc en la quantificació d'incertesa:

- Càlcul de mètriques com la desviació estàndard, l'RMSE i el Spread-to-Error Ratio (SER).

- Avaluació de la dispersió espacial i temporal dels ensembles.

- Disseny de figures i visualitzacions per a l'anàlisi global i regional dels resultats.

- Redacció completa de la memòria, incloent l'estructura metodològica, la descripció tècnica del workflow, la discussió dels resultats i les limitacions detectades.

- Proposta de línies futures com la integració de nous models, la millora del downscaling i l'ús de dades reals per a la validació.

- Presentació d'una versió preliminar del treball al Doctoral Symposium del BSC.

Relació del treball que es presenta amb els Objectius de Desenvolupament Sostenible.

El treball presentat s'emmarca dins l'àmbit de la predicció climàtica i contribueix directament a diversos Objectius de Desenvolupament Sostenible formulats per les Nacions Unides, especialment en el context de la lluita contra el canvi climàtic i el foment d'una ciència oberta, eficient i orientada a l'impacte social.

En primer lloc, el projecte s'alinea clarament amb l'**ODS 13: Acció pel Clima**, que promou accions urgents per combatre el canvi climàtic i els seus efectes. L'objectiu central del treball és millorar la quantificació de la incertesa en les prediccions climàtiques d'alta resolució, proporcionant eines més eficients per a l'adaptació i mitigació dels impactes climàtics a escala regional. Això és especialment rellevant per a sectors vulnerables com l'agricultura, la gestió de recursos hídrics o la planificació urbana.

A més, la metodologia desenvolupada en aquest treball contribueix a fer més accessibles els sistemes de modelització climàtica a institucions i regions amb recursos computacionals limitats. En aquest sentit, dona suport a l'**ODS 9: Indústria, Innovació i Infraestructura**, promovent la innovació tecnològica i la transferència de coneixement mitjançant workflows modulars, contenitzats i reutilitzables.

La integració d'intel·ligència artificial com a alternativa als models físics tradicionals contribueix també a l'**ODS 12: Producció i consum responsables**, en tant que redueix dràsticament el consum energètic i de recursos computacionals associats a les simulacions climàtiques. Aquest enfoc permet generar prediccions útils amb una petjada de carboni molt menor, fet essencial per al desenvolupament de pràctiques científiques sostenibles.

Finalment, el projecte promou principis de ciència oberta, interoperabilitat i reutilització (principis FAIR), contribuint indirectament a l'**ODS 17: Aliances per assolir els objectius**, en tant que facilita la col·laboració entre institucions, la replicabilitat dels resultats i la creació d'eines compartides per a la comunitat científica global.

FORMATTED REPORT

AI-Assisted Uncertainty Quantification in High-Resolution Climate Predictions

Author: Aina Gaya Àvila
*Earth Sciences Department, Barcelona Supercomputing Center (BSC), Plaça d'Eusebi Güell, 1-3, Les Corts, 08034 Barcelona**

Advisor: Oriol Tintó Prims

Abstract: Strategies to adapt and mitigate climate change demand global high-resolution information. The initiatives producing this information currently lack uncertainty estimates, since traditional approaches for uncertainty quantification, which rely on running multiple simulations with slightly different parameters or initial conditions, are computationally expensive. This study presents an alternative to this problem: a framework to quantify uncertainty using Artificial Intelligence-based models, which are cheaper to run. We present a modular framework that successfully initializes two AI-based models, NeuralGCM and AIFS, with high-resolution climate data from the Destination Earth Climate Digital Twin. Using this framework under warmer climate conditions, we reproduce the systematic cold drift with AIFS presented in previous studies, while NeuralGCM maintains physical consistency with physics-based simulations. NeuralGCM generates consistent ensemble spreads from which we can estimate the uncertainty of the high-resolution climate data. NeuralGCM ensembles required only 1% of the cost of equivalent physics-based simulations, demonstrating the framework's potential as a computationally viable alternative for high-resolution climate uncertainty quantification.

I. Introduction

Climate change is driving profound and challenging transformations in ecosystems, societies, and economies worldwide. Understanding and anticipating the implications of these changes requires accurate and actionable climate information at relevant spatial and temporal scales [1]. Global circulation models (GCMs) offer important insights into climate patterns [2], but local decision-making, particularly in the context of extreme events such as floods, droughts, or heatwaves, demands high-resolution, region-specific projections [3–6].

This need for localized climate information is particularly acute in sectors such as agriculture, energy, urban planning, and disaster management. In these areas, planning decisions must consider fine-scale environmental variations. Traditional GCMs often fall short in resolving those variations because they can not resolve critical physical processes like ocean eddies or convective storms due to computational limitations [7]. Recognizing this gap, the scientific community has made significant investments in high-resolution modelling capabilities [8, 9], using the advances in computational power to develop more sophisticated simulations of the Earth system.

A. Destination Earth - Climate Adaptation Digital Twin

Among the most ambitious initiatives addressing this challenge is Destination Earth (DestinE) [10], a flagship program by the European Commission designed to develop accurate digital twins of the Earth. These digital twins aim to transform the way climate data is generated and consumed, directly enabling scientists, policymakers, and stakeholders to explore *what-if* scenarios and support evidence-based decision-making across multiple scales. At the heart of this vision is the Climate Change Adaptation Digital Twin (ClimateDT) [11]. It aims to provide operational global climate simulations at an unprecedented 5 km scale. Those simulations run on supercomputers deployed by EuroHPC JU, a joint initiative between the European Union, European countries, and private partners to develop a world-class supercomputing ecosystem. The simulations are currently running in MareNostrum5 (Barcelona) and LUMI (Finland).

The digital twin generates data running physics-based models, processes it online, and provides it to real-world applications, which are integrated into the operational system, to produce indicators.

However, producing reliable uncertainty estimates for such high-resolution projections remains a major challenge [12], even though it is an essential component of climate prediction given the chaotic nature of the system. The traditional way of producing uncertainty estimates for lower resolution GCMs, used by most of the operational weather prediction facilities worldwide, is using **ensembles**. Ensemble forecasting consists of running multiple simulations with slightly varied initial conditions or model parameters [13, 14], which lead to different trajectories. Those multiple simulations are conducted to account for the two usual sources of uncertainty in weather forecasting: the fact that the initial conditions are not perfect, which is amplified by the chaotic nature of the equations that govern the atmosphere, and the errors introduced by the models, such as the approximations made to solve those equations.

Producing ensembles is computationally expensive and often unfeasible at the resolutions targeted by DestinE, since at this level of detail, there are big challenges in

*Electronic address: agayaavi53@alumnes.ub.edu

both computation and data storage. Running a single ClimateDT simulation for 50 simulated years requires approximately 100 days of continuous computation on 200 nodes, equivalent to 480,000 node-hours. At current pricing (0.57 EUR per node-hour on LUMI-C [15]), a single simulation costs 273,600 EUR. A typical 50-member ensemble, therefore, could exceed 10 million EUR. Those same simulations produce about 80 GB per simulated day, which adds up 1.5 PB for a 50-year simulation. This creates a significant gap in the ability to quantify confidence and risk in future high-resolution climate projections.

B. Alternative to physics-based simulations: AI-based models

An alternative to address these computational challenges can potentially be found in models based on artificial intelligence (AI). Those models are fed with large amounts of weather and climate data from the past, from which they learn patterns. Once trained, they are able to predict what comes next based on those patterns. Recent advances in AI-based models have shown the capability to generate forecasts at significantly reduced computational costs [16–18]. Several of these models now perform on par with, or even outperform, traditional numerical weather prediction systems in specific benchmarks [19].

A key emerging application is the prediction of extreme events, where AI models are increasingly being explored for their potential [20, 21]. Nevertheless, the performance of those models in climate prediction remains an open question. Recent studies running AI models in climates different from their training conditions have revealed systematic biases, such as a cold drift towards the state where the model was trained (typically based on reanalysis data) when initialized in warmer climate regimes [22].

AI models for weather and climate science have emerged as one of the most rapidly expanding research frontiers. Given this pace of innovation, where new models and methodologies emerge monthly, any system developed for weather and climate applications must be designed with modularity as a core principle, employing containerized components, and flexible data pipelines that can integrate new models, adapt to evolving datasets, and scale with the field's rapidly changing computational paradigms without requiring complete system refactoring.

C. The need for efficient workflows in the climate community

One of the main challenges when working with climate simulations is the volume and complexity of the data involved. The climate community relies on a variety of formats and storage solutions, which makes handling and transforming datasets a demanding process, especially at large scales. This heterogeneity, combined with the need for fast and reproducible analysis, highlights the importance of developing efficient workflows. For example, the current operational runs in ClimateDT contain more than 130.000 jobs. To align with the FAIR (Findable, Accessible, Interoperable, Reusable) principles [23], research institutions are taking significant steps to develop portable and configurable workflows. These workflows are shared among scientists, integrate various models, support ensemble generation, and allow for monitoring and diagnostics across High-Performance Computing (HPC) platforms. As a result, researchers can focus on their scientific questions instead of the complexities of HPC.

The integration of AI into high-resolution climate modeling has similar workflow requirements. Traditional Earth System Models (ESM) require robust tools to manage large-scale simulations across HPC environments, and AI-based models demand similar efficient, standardized workflows to ensure scalability, reproducibility, and ease of use. Workflow managers such as Autosubmit [24] play a key role in this challenge by orchestrating complex, large, and multi-step simulation tasks across heterogeneous computing environments, and also help to make the simulations reproducible and to ensure proper provenance tracking by systematically logging configurations, inputs, outputs, and execution metadata.

D. Objectives

The primary aim of this study is to build a modular, end-to-end framework that integrates AI-based forecasting with high-resolution climate projections to assess uncertainty through ensemble prediction.

At the same time, we explore whether AI-based models could be an alternative to overcome the limitations of traditional ensemble simulations in the context of ClimateDT, which is considered state-of-the-art in climate sciences.

By applying this framework in different use cases, we determine whether AI-generated ensembles can deliver uncertainty estimates, potentially overcoming the computational constraints of classical ensemble simulations in the context of high-resolution climate projections. Once set, researchers can use it for multiple scientific experiments, such as the ones shown in Section III.

The framework has three steps: (1) generating input data for the AI models using the output of high-resolution physical models, (2) generating AI-driven ensembles with multiple AI models, and (3) evaluating ensemble spread and providing uncertainty quantification metrics.

II. Methodology

Several design decisions have to be made in order to build a framework such as the one described. The models and the sources of data need to be chosen, the software has to be designed, and the metrics to quantify the uncertainty need to be established. In the following sections, we describe our design decisions.

A. Description of the framework

In order to run ensemble AI-based simulations starting from ClimateDT data, one has to perform a series of steps. Those are:

- Data preparation
- Simulation
- Post-process
- Visualization and interpretation

If, additionally, one wants to run multiple simulations (starting from a diverse range of initial dates, or performing multi-member runs), then the need for a workflow and a workflow manager becomes undeniable.

The proposed framework is model-agnostic and modular. This allows the integration of different AI-based forecasting models and the addition of other sources of initial conditions.

The code structure facilitates the addition of new models. It uses a list of variables the AI model needs, the format the model accepts, and the usual interface for interaction, which is often a Python call. The code is stored in `GitHub` [38].

The workflow uses Autosubmit, the workflow manager developed in the Earth Sciences Department at the Barcelona Supercomputing Center, which supports different ESM and multiple Digital Twin initiatives (such as DestinE). It helps us to organise, set up, and run long simulations systematically. It is well suited to our workflow, since it is built for climate simulations, so we can set up jobs with a calendar structure.

The structure of any workflow using Autosubmit includes configuration files, which define the parameters used in an experiment, templates, scripts containing placeholders that are filled with the parameters defined in the configuration files, and supporting software.

1. *Workflow configuration*

To set up a simulation, the user must define a range of parameters. These parameters encompass: the starting date and duration of the simulation, the model to be executed, the number of ensemble members, and the location of the initial data, among others. This configuration is made accessible to the user via a YAML file, and is the experiment-specific configuration.

The rest of the configuration is concealed from the user and is stored within the `conf` directory of the git repository. Based on the user's selections, a specific subset of the configuration is utilized.

2. *Templates and runscripts*

The `templates`, together with the configuration, build the scripts that Autosubmit submits to the scheduler of the HPC. These template files contain placeholders. Before Autosubmit sends these templates to the scheduler, it reads the parameters from the configuration files and replaces each placeholder in the template with the actual value specified in the configuration. This process creates customized scripts tailored to the current experiment, which are submitted to the HPC as jobs.

In the `runscripts` directory, a collection of functional auxiliary Python scripts is stored. These scripts are not meant to be run directly by the user; instead, they are designed to perform specific tasks or computations that support the main workflow. This modular approach helps keep the main configuration and templates clean and organized while enabling reusable and maintainable code for common tasks.

3. *Containers*

HPC environments vary significantly in architecture, libraries, and compilers, creating portability challenges across different systems. To address this, we developed the entire simulation pipeline using Singularity containers. A container is a lightweight, standalone software package that encapsulates an application along with all its dependencies, libraries, runtime configurations, and system tools into a single, portable unit. This approach isolates the software from the host system, making the deployment easier and ensuring that our models execute identically when running on different HPC clusters.

This containerization strategy enables seamless collaboration between institutions and eliminates the common *it works on my machine* problem. More importantly, it ensures long-term reproducibility by preserving the exact computational environment used for each simulation. Researchers can recreate identical conditions months or years later, allowing for scientific validation, independent verification, and reliable building upon previous results.

B. Stochastic vs deterministic AI models

The atmosphere's inherent chaotic nature implies that tiny differences in initial conditions can lead to vastly different outcomes: the butterfly effect. Traditional weather and climate models address this by running ensembles with slightly different starting points to capture uncertainty. Can we use the same approach for AI models?

1. *Deterministic models*

Deterministic AI models produce the same output for a given input every time, with no inherent randomness in their predictions [25]. In climate modeling, deterministic AI models are often trained to minimize prediction error using loss functions such as root mean squared error (RMSE), aiming for accuracy and stability over long-term simulations. The first generations of AI models, which are trained to minimize RMSE, tend not to show this effect [26], so we cannot use them to generate ensembles. The next generation of deterministic models could potentially be able to reproduce this effect.

2. *Stochastic models*

Stochastic AI models incorporate randomness, either in their architecture or during prediction, allowing them to

generate multiple plausible future scenarios for the same input. In climate modeling, stochastic AI models may inject random noise into neural network components or use loss functions (e.g., Continuous Ranked Probability Score, CRPS) that encourage ensemble spread, representing the uncertainty and inherent variability of the climate system.

C. Data-driven models used

Two state-of-the-art AI-based forecasting systems were selected for this work: **NeuralGCM** [27] and **AIFS** [28]. Both models can be easily used because the user can download pre-trained, ready-to-use models, and both are well-documented for direct usage. In this project, we are using those models and running inference with them on new data, but it is worth mentioning that we are not retraining or fine-tuning the models.

The models that we use are designed to run on GPUs, so we deployed the models and ran them in the accelerated partition (GPU) of MareNostrum5 [29].

1. NeuralGCM model

NeuralGCM is a differentiable hybrid atmospheric model developed by Google Research and the European Center for Medium-Range Weather Forecasts (ECMWF). Similar to conventional models, NeuralGCM divides the Earth's atmosphere into cubic units and performs calculations on the physics governing large-scale phenomena such as the movement of air and moisture. However, rather than relying on parameterizations[39], it employs a neural network to derive the physics of these events from historical weather data.

Since we were interested in stochastic models, through this work, we used the stochastic pre-trained model provided by NeuralGCM, which has a 1.4-degree resolution (about 100km) and 37 pressure levels.

2. AIFS model

Artificial Intelligence Forecasting System (AIFS) is a model developed by ECMWF. Since February 2025, it has been considered operational, providing forecast predictions every 6 hours. They report a reduction of approximately 1,000 times in energy use for making a forecast, compared to the physics-based model of the same center, IFS [28].

The output of the model is in 0.25 degrees (about 36 km) and in 13 pressure levels.

AIFS-single, the one currently open-source and used throughout this work, is a deterministic model. There is a stochastic version of the model [30] which is not yet open-source, but would be used to produce ensembles.

D. Initial conditions preparation

In this work, we are considering mainly two sources of data, but the modular approach makes it potentially expandable.

The first is ERA5, which is the fifth generation of ECMWF's atmospheric reanalysis dataset and is considered the best representation of the atmosphere through the years. ERA5 provides hourly estimates of atmospheric, land, and oceanic climate variables by combining vast amounts of historical observations with numerical weather prediction models through data assimilation techniques. This makes it one of the most complete and consistent global climate datasets available, offering unprecedented detail and accuracy for climate research and machine learning applications. ERA5 covers the period from 1940 to the present and is the dataset that most AI-based models use for training. The data covers the entire Earth on a 31km grid and resolves the atmosphere using 137 levels from the surface up to a height of 80km. The dataset can be accessed via the Copernicus Climate Data Store [31].

The second and main source of interest in this work is ClimateDT data, which has 5km resolution in the atmosphere and 10km resolution in the ocean. It is temporarily stored in a database in MareNostrum5, where the simulations have run.

AIFS and NeuralGCM have different input data requirements. They need different variables as inputs, in different grids and levels, and with different metadata. Luckily, most of the variables were produced by the Numerical Models, but one of them was not (ciwc, see Table I), and has been inferred (Sec. II D 1).

The input requirements for 3D variables among different models are also different: DestinE outputs 3D variables in 19 pressure levels, AIFS asks for a subset of these pressure levels, while NeuralGCM demands more pressure levels than those provided by DestinE (37). To solve this, we are interpolating to the levels that are not provided.

1. Inference of the missing variables

One variable that was not provided by the GCMs was the specific cloud ice water content (ciwc), required by NeuralGCM. ciwc is a 3D variable that measures the mass of cloud ice per kilogram of moist air (kg/kg).

Cloud water can be liquid, ice, or both. To estimate ciwc, we used 3 variables outputted by ClimateDT:

- Total column cloud ice water (tciw, 2D): total ice in clouds from the surface to the top of the atmosphere (kg/m^2).
- Specific cloud liquid water content (clwc, 3D): cloud liquid per kilogram of moist air (kg/kg).
- Total column cloud liquid water (tclw, 2D): total liquid water in clouds in a vertical column (kg/m^2). It excludes rain.

Our method was to scale the clwc profile by the ratio tciw / tclw. However, this assumes a similar vertical distribution for cloud liquid and ice, which is not physically accurate, since liquid tends to be lower in the atmosphere (lower to middle troposphere), while ice is found higher up (middle to upper troposphere).

Because this estimate produced values that differed significantly from the model's training data and caused

Variable	Required by NeuralGCM	Required by AIFS	Available in DestinE
u_component_of_wind	✓	✓	✓
v_component_of_wind	✓	✓	✓
geopotential	✓	✓	✓
temperature	✓	✓	✓
specific_humidity	✓	✓	✓
vertical_velocity	✗	✓	✓
specific_cloud_liquid_water_content	✓	✗	✓
specific_cloud_ice_water_content	✓	✗	✗
sea_surface_temperature	✓	✓	✓
sea_ice_cover	✓	✗	✓
surface_pressure	✗	✓	✓
mean_sea_level_pressure	✗	✓	✓
2m_temperature	✗	✓	✓
2m_dewpoint_temperature	✗	✓	✓
10m_u_component_of_wind	✗	✓	✓
10m_v_component_of_wind	✗	✓	✓
land_sea_mask	✗	✓	✓
orography	✗	✓	✓
std_dev_subgrid_orography	✗	✓	✓
slope_subgrid_orography	✗	✓	✓
insolation	✗	✓	✓
latitude/longitude	✓	✓	✓
time_of_day / day_of_year	✓	✓	✓

TABLE I: Comparison of variable requirements between NeuralGCM and AIFS models versus availability in DestinE ClimateDT output, highlighting the need for a variable inference (ciwc). ✓ means that the variable is required/available, while ✗ means that the variable is not required by the model or not available in DestinE.

instabilities, we added an empirical scaling factor of 1/400. This solution allowed us to run simulations, although the impact of this approximation has not been fully studied. We recommend including ciwc as an output variable in future iterations of ClimateDT to avoid these issues.

E. Uncertainty Analysis

Stochastic models allow us to generate ensembles in which each member evolves slightly differently due to random perturbations. The extent to which ensemble members diverge from one another or from a reference trajectory provides information about the system's uncertainty at a given point in time.

To quantify this uncertainty, we compute the Spread, and to assess the performance of the AI models against the physics-based simulation, the Root Mean Square Error (RMSE) and the Spread-to-Error Ratio (SER).

1. Root Mean Square Error (RMSE)

The RMSE is a widely used statistical metric that quantifies the average magnitude of the difference between predicted values and a reference. At a given point in space and time, it is defined as:

$$\text{RMSE} = \|\hat{x} - x\| \tag{1}$$

where $\hat{x}_i$ is the reference value at grid point i, x_i is the predicted value. Averaging in time or space, it is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - x_i)^2} \tag{2}$$

where N is the total number of points.

2. Spread

The spread (σ_{ensemble}) is defined as the square root of the mean ensemble variance [32], and it is a measure of the differences between the ensemble members. It is defined as,

$$\sigma^2_{\text{ensemble}} = \frac{1}{M}\sum_{i=1}^{N}(x_i - \hat{x})^2 \tag{3}$$

Where M is the number of ensemble members and $\hat{x}$ is the ensemble mean.

A small spread indicates lower uncertainty, while a large spread indicates higher uncertainty.

3. Spread-to-Error Ratio (SER)

The Spread-to-Error Ratio (SER) compares the ensemble spread to the RMSE and provides insight into the ensemble's reliability. The SER is defined as:

$$\text{SER} = \frac{\sigma_{\text{ensemble}}}{\text{RMSE}} \tag{4}$$

where σ_{ensemble} is the standard deviation across ensemble members at each grid point (averaged over space or time, depending on context).

A perfectly reliable forecast has $SER = 1$, indicating a balanced spread and error. Deviations from this ideal reveal systematic biases in uncertainty representation. A spread/error ratio greater than one indicates an underconfident ensemble, whereas a spread/error ratio less than one indicates an overconfident ensemble. The models are usually tuned to produce a SER close to 1.

III. Results and discussion

In all the use-cases shown below, a 20-year (2020-2040) simulation performed with IFS-FESOM [33] model has been used as reference data and to create the initial conditions. IFS-FESOM is one of the three coupled models used in ClimateDT, and simulates both the atmosphere and the ocean. This simulation uses the SSP3-7.0 scenario [34] and predicts a warmer world in 2039 (Fig. 1).

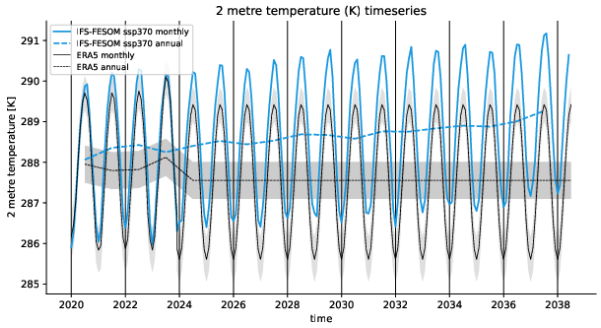

FIG. 1: Two-meter temperature from the ClimateDT IFS-FESOM coupled model simulation for 2039, representing the warmer climate conditions (SSP3-7.0 scenario) used as initial conditions and reference data for AI model evaluation. Source: AQUA [35].

The main variable under study through all the use-cases has been the temperature at 1000 hPa, near the surface.

A. Use-case 1: running 10 days in a warmer climate

For this first use case, the goal was to check whether the outputs of the AI models were able to follow similar trajectories as the physics-based models, if the models under consideration showed a cold drift towards the training state, and if the model was performing similarly over the globe.

Each model was configured to produce 10-day forecasts starting from initial conditions generated for January 2039. Both AIFS and NeuralGCM were configured to run for 10 days, producing outputs every six hours.

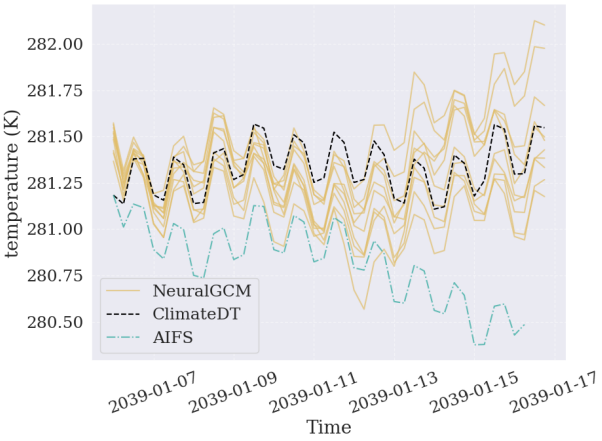

FIG. 2: Global mean temperature at 1000 hPa over 10 days comparing ClimateDT reference data (physics-based), AIFS deterministic forecast, and NeuralGCM 10-member ensemble initialized from January 2039 conditions.

1. *Global analysis*

The timeseries for temperature for both NeuralGCM (10-member ensemble) and AIFS (single member) can be seen in Fig. 2. It can be seen how AIFS drifts to a colder state, more similar to where it was trained. These results are consistent with the ones in [22]. On the other hand, the different ensemble members of NeuralGCM tend to follow a trajectory similar to the one described in ClimateDT.

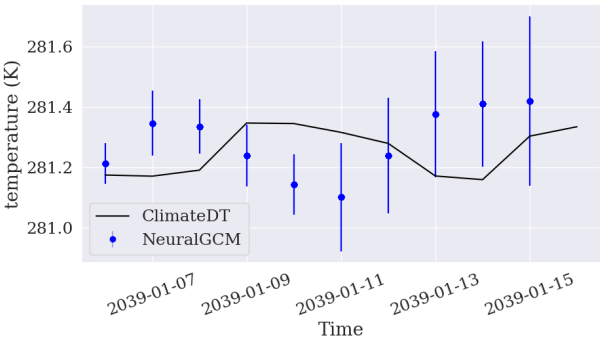

FIG. 3: Global mean temperature at 1000 hPa showing NeuralGCM ensemble mean with uncertainty bounds (error bars) calculated with the spread, compared to ClimateDT reference trajectory over 10 days (Use-case I).

In Fig. 3, we show the daily mean and the spread given by the different members in the ensemble, and we can see how the error bars do not entirely cover the trajectory of the GCM, but do not deviate substantially. We can say, given these results, that NeuralGCM is giving trajectories that are consistent with those given by the physics-based models, even in a warmer climate, as opposed to

the results previously published [22], which involve a cold drift in AIFS, Graphcast, and Panguweather.

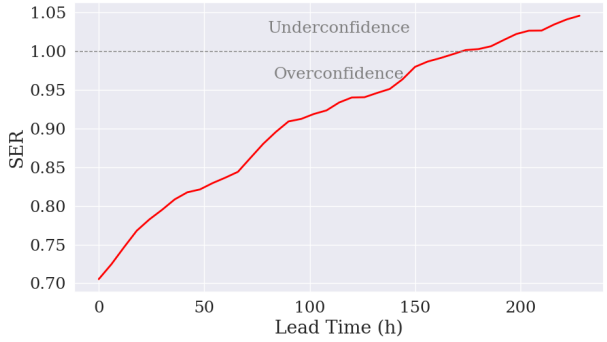

FIG. 4: Spread-to-Error Ratio (SER) evolution for NeuralGCM temperature forecasts at 1000 hPa, over the 10-day period.

Fig. 4 shows the evolution of the Spread to Error Ratio (SER) over time. The SER is in the "Overconfidence" area for the first days of simulation, but as time passes, it evolves towards the ideal value (SER = 1), until it is surpassed and it enters the "Underconfidence" area. This means that during most of the simulation, there is more error than spread, but towards the end of the simulation, the error estimates are more realistic than in the beginning, and even overestimated.

2. Spatial analysis

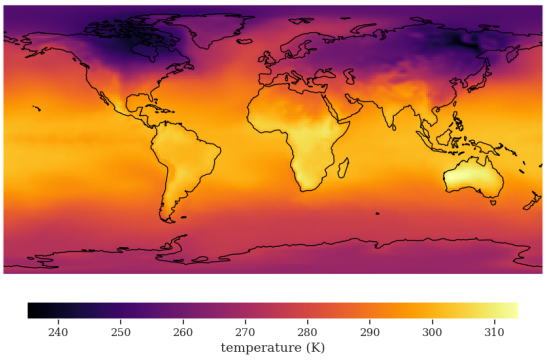

FIG. 5: Global temperature at 1000 hPa averaged across 10 NeuralGCM ensemble members averaged over the complete 10-day simulation period, showing spatially coherent patterns consistent with physics-based simulations.

Averaging over time the 10 ensemble members of NeuralGCM, we see maps consistent with observations for the 1000 hPa temperature over the globe (Fig. 5). Plotting the standard deviation (Fig. 6, we see more variability in the northern high latitudes (Siberia, Alaska, Northern Canada), Antarctica, and southern oceans. Those can be considered the areas with more uncertainty for the period under study. At short time-scales, atmospheric dynamics are the main driver of uncertainty: the positions of low and high pressure systems in the initial conditions largely determine which regions are associated with higher forecast uncertainty.

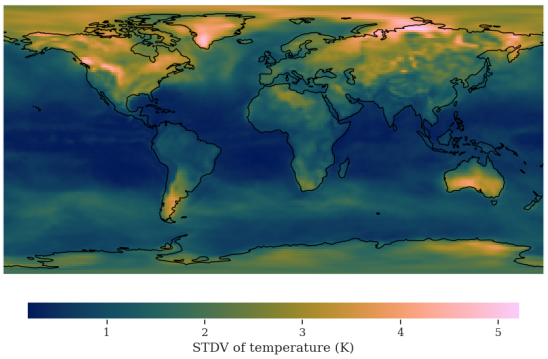

FIG. 6: Standard deviation of temperature at 1000 hPa across 10 NeuralGCM ensemble members, averaged over the 10-day simulation period.

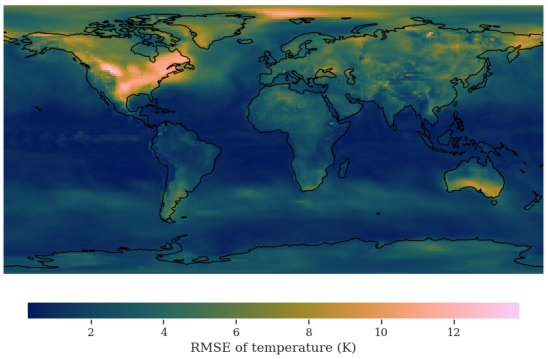

FIG. 7: Root Mean Square Error (RMSE) of NeuralGCM temperature at 1000 hPa relative to ClimateDT reference, with darker regions indicating higher forecast skill and lighter areas showing reduced accuracy.

The spatial distribution of prediction errors is shown in Fig. 7, where NeuralGCM demonstrates higher accuracy in regions with lower RMSE values (darker areas) and reduced skill in regions with higher errors (lighter areas). The spread-to-error ratio analysis in Fig. 8 reveals the calibration quality of the ensemble predictions across different regions. For a perfectly calibrated ensemble, the SER should equal unity ($\log(\text{SER}) = 0$) globally, indicating that the ensemble spread accurately represents the prediction uncertainty. In the darkest blue areas, the model is underdispersive, meaning that the model is reporting less uncertainty than it should. In the red areas, the model is overdispersive, meaning that it is estimating more uncertainty than it should.

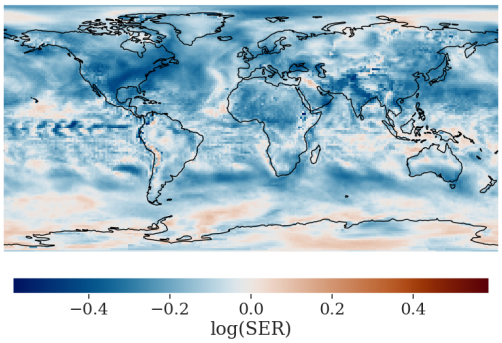

FIG. 8: Spatial distribution of Spread-to-Error Ratio (SER) for NeuralGCM temperature forecasts, averaged over 10 days, where blue regions indicate underdispersive behavior (underestimated uncertainty) and red regions show overdispersive behavior (overestimated uncertainty).

B. Use-case 2: running one month in a warmer climate

In this second use-case the models are initialized every day of January 2039, and run for 10 days.

1. *AIFS analysis*

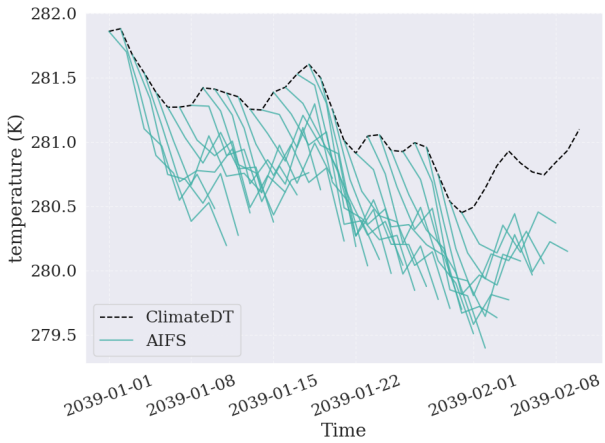

FIG. 9: Global mean temperature at 1000 hPa from AIFS simulations initialized daily throughout January 2039, demonstrating systematic cooling toward training state conditions over 10-day forecasts.

The results presented in Fig. 9 show a clear tendency of the model to cool down to the initial state, which is consistent with the results presented in [22].

2. *NeuralGCM analysis*

We ran a 10-member ensemble using NeuralGCM, with each member initialized daily from ClimateDT data throughout January 2039. The evolution of these ensemble members is shown in Fig. 10.

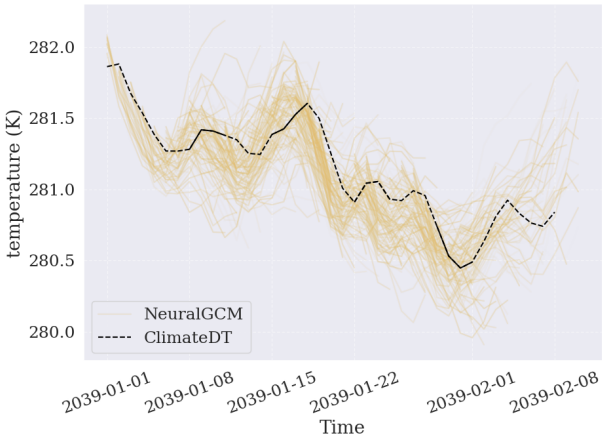

FIG. 10: Global mean temperature at 1000 hPa from NeuralGCM 10-member ensembles initialized daily throughout January 2039, showing consistent evolution and reduced cold drift compared to AIFS.

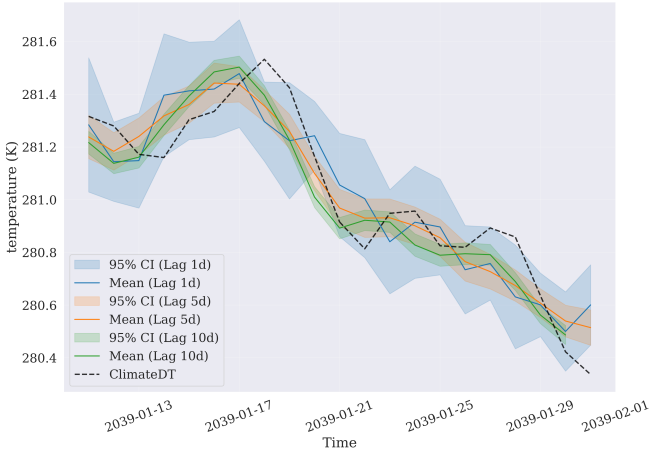

FIG. 11: Global mean temperature at 1000 hPa with 95% confidence intervals derived from NeuralGCM lagged ensembles, constructed by aggregating forecasts from multiple daily initializations throughout January 2039.

NeuralGCM simulations exhibit a consistent evolution across ensemble members, capturing the main trends in global temperature at 1000 hPa. Compared to AIFS, NeuralGCM displays a reduced tendency to revert to the initial state. The spread across ensemble members grows over forecast lead time, reflecting the model's internal variability and uncertainty in the initial conditions. This behavior highlights the model's ability to maintain physical consistency while still generating a meaningful ensemble spread.

Fig. 11 shows the confidence intervals for lagged forecasts. To construct these confidence intervals, we generated lagged ensembles by aggregating forecast outputs from multiple initialization times. For each forecast lead time, we computed the sample mean $\hat{x}$, the standard de-

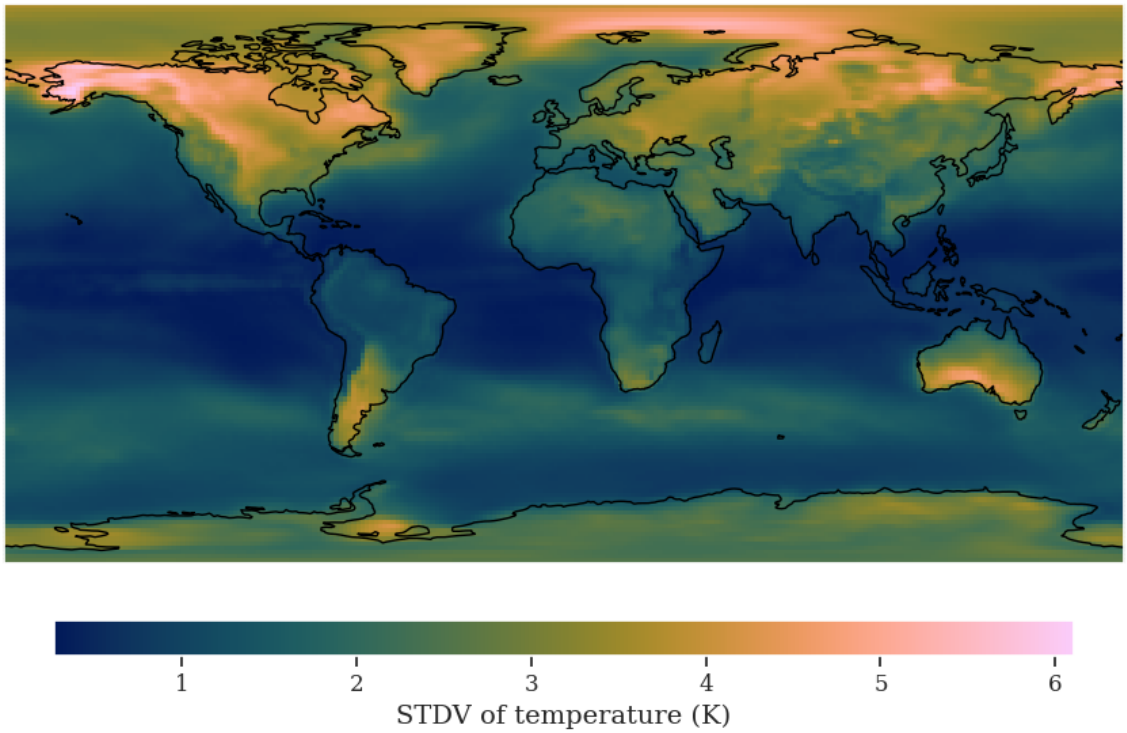

FIG. 12: Average standard deviation of temperature at 1000 hPa from NeuralGCM daily-initialized ensembles for January 2039, showing larger uncertainty in the Northern Hemisphere and demonstrating the model's ability to capture regional variability in forecast confidence.

viation σ, and the number of ensemble members M. Assuming approximately normally distributed forecast values, the confidence intervals were computed using the standard formula:

$$\left[\hat{x} - \frac{c\sigma}{\sqrt{M}}, \hat{x} + \frac{c\sigma}{\sqrt{M}}\right], \tag{5}$$

where c is the critical value from the standard normal distribution corresponding to the desired confidence level (e.g., $c = 1.96$ for a 95% confidence interval).

3. *Spatial analysis*

From the spatial analysis, it can be seen how the spread is more prominent in the Northern Hemisphere (Fig. 12). This means that the model tends to diverge more in those areas, predicting higher uncertainty. The RMSE (Fig. 13) is similar to the previous use case. The SER (Fig. 14) tends to be more balanced, giving better predictions, probably because we are averaging over more data and therefore, the initial conditions play a less crucial role.

C. Performance analysis

1. *NeuralGCM*

In order to produce use-case 2, 310 jobs were submitted (10 ensemble members per day). The mean execution

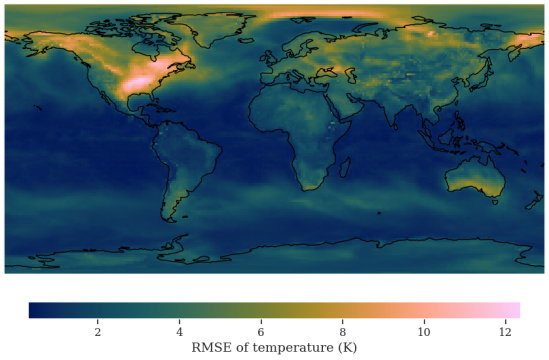

FIG. 13: Root Mean Square Error (RMSE) of temperature at 1000 hPa for NeuralGCM ensembles with daily initialization for January 2039, showing spatial patterns of model skill consistent with those presented in Use-case I.

time per job was 85.13 ± 7.28 seconds per job, which adds up to 7.33 node hours. An approximation to extend it for the 50 simulated years in the ClimateDT protocol adds up to about 4.300 GPU node hours, 1% of the cost of running the physics-based simulations.

We compared our performance in simulated years per day (SYPD) with the one reported in [27], and saw that

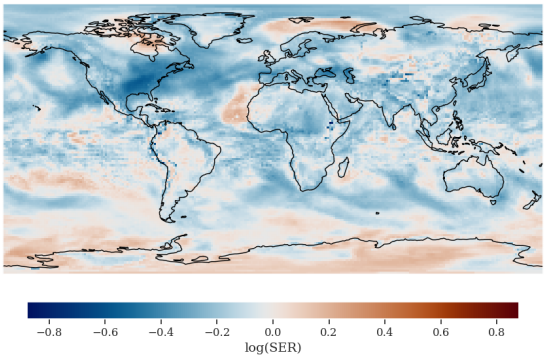

FIG. 14: Spread-to-Error Ratio (SER) for temperature at 1000 hPa from daily-initialized NeuralGCM ensembles for January 2039, showing a more balanced SER compared to single-initialization experiments.

it was substantially slower in our case. Isolating the runtime of the model (taking off any post-processing), we ran simulations of different lengths, ranging from 1 to 31 days, and detected an initialization time of 29.17 seconds, while simulating a day takes 2.31 seconds. This makes running longer simulations more efficient, as can be seen in Fig. 15. Extrapolating to longer integration times gives us results comparable to the ones reported by NeuralGCM's team [27].

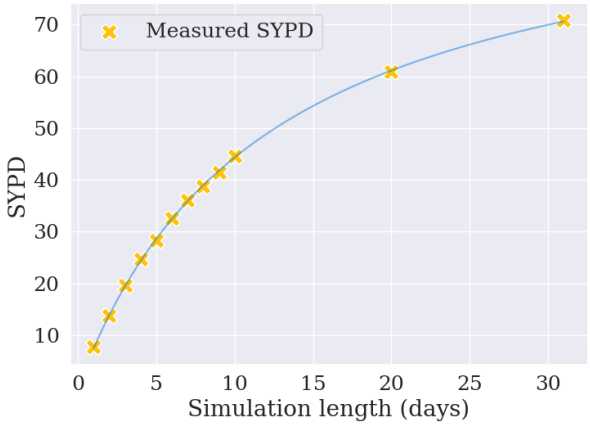

FIG. 15: NeuralGCM simulated years days per day (SYPD) versus forecast length, showing decreasing computational cost per simulated day due to the amortization of 29.17-second initialization overhead, with optimal efficiency achieved for longer integrations.

We used EAR [36] to measure the energy consumed by our jobs. In terms of energy consumption, the impact of the forecast length decreases exponentially, as can be seen in Fig. 16. The results indicate a strong initial decrease in energy per simulated year as the number of simulated days increases, suggesting improved computational efficiency for longer simulations due to the amortization of the initialization costs, confirming the results from the SYPD analysis.

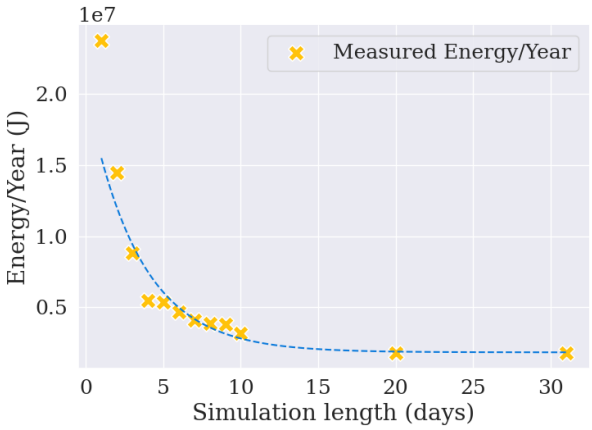

FIG. 16: Energy consumption per simulated year versus forecast length for NeuralGCM, demonstrating exponential decrease in energy costs for longer simulations due to initialization overhead amortization.

2. AIFS

For AIFS, in use-case 2, we spent 113.35 ± 16.37 seconds/job (10 days). Only 31 jobs ran because the model is deterministic, and therefore, we are not producing an ensemble with it. ECMWF reports 250 seconds/15-days with AIFS-Diffusion [37], meaning that our performance results are comparable to, or even better than, the ones reported by the AIFS team.

D. Limitations and challenges

Despite the insights provided by this analysis, several limitations should be acknowledged. The resolution mismatch between AI models and physics-based simulations presents another significant challenge. While ClimateDT provides 5km resolution data, the AI models operate at considerably coarser resolutions (100km for NeuralGCM, 36 km for AIFS), potentially limiting their utility for local-scale applications that motivated this research. Furthermore, the reference dataset used for comparison is not based on observations but rather on output from a separate numerical simulation, which limits the ability to assess the absolute realism of the models and confines the evaluation to a relative comparison. Beyond this data limitation, all models employed in this study are subject to structural uncertainties and may exhibit systematic biases that can affect the fidelity of forecasts and the reliability of uncertainty quantification metrics derived from them. Compounding these model-related issues, the ensemble size for each model is relatively small (e.g., 10 members), which may not be sufficient to robustly capture the full range of internal variability or to derive statistically significant uncertainty estimates.

Finally, the preprocessing steps required to harmonize model outputs, including temporal alignment, spatial interpolation, and variable selection, introduce additional sources of uncertainty that may influence the results.

IV. Conclusions

This research successfully demonstrates the feasibility of AI-assisted uncertainty quantification for high-resolution climate projections, presenting a framework that bridges the gap between computationally expensive traditional ensemble methods and the need for practical uncertainty estimates.

The developed framework provides a model-agnostic, modular system that can integrate various AI-based forecasting models with high-resolution climate data. The successful deployment of both NeuralGCM and AIFS models within this framework demonstrates its versatility and potential for broader application across the climate modeling community.

The containerized workflow architecture ensures reproducibility and portability across different computing environments, addressing a critical need in climate science for standardized and FAIR-compliant research practices. The integration with Autosubmit is a step towards the potential extensive use of the framework among the community within existing climate modeling infrastructures.

Using this framework, we were able to easily reproduce the results from [22]: AIFS presents a systematic drift towards its training state when initialized in warmer conditions. Opposite to this, we have seen how NeuralGCM is able to produce physically consistent results that follow the trajectory of the original physics-based simulation.

Finally, we were able to compute uncertainty quantification metrics (σ, RMSE, and SER) and evaluate in which regions the spread and the error are more prominent.

This work is just the basis of a set of potential uses of AI-based climate models. Future work encompasses several key areas, beginning with enhancements to the framework itself. First, model integration represents a crucial next step, involving the incorporation of other state-of-the-art models such as AIFS-CRPS, PanguWeather, or GraphCast to expand the analysis. In addition to expanding the model suite, incorporating other sources of data to initialize the models could be beneficial for certain analyses. Moreover, addressing the resolution gap between AI models and high-resolution climate data necessitates the development of downscaling techniques to provide local-scale climate features essential for impact assessment.

Beyond these framework improvements, scientific validation presents equally important opportunities for advancement. Specifically, comprehensive model validation through historical simulations would enable assessment of model performance against observational data, providing crucial insights into absolute model accuracy rather than relative comparisons. Furthermore, expanding the temporal scope of analysis beyond the data from January 2039 processed in this study would be valuable, as examining seasonal variations through analysis of different periods of the year could reveal important seasonal dependencies in model performance and uncertainty characteristics.

[1] H. Lee *et al.*, "IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change," tech. rep., Intergovernmental Panel on Climate Change (IPCC), July 2023.

[2] V. Eyring *et al.*, "Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization," *Geoscientific Model Development*, vol. 9, no. 5, pp. 1937–1958, 2016.

[3] T. Rackow *et al.*, "Delayed Antarctic sea-ice decline in high-resolution climate change simulations," *Nature Communications*, vol. 13, p. 637, Feb. 2022.

[4] W. J. Gutowski *et al.*, "The Ongoing Need for High-Resolution Regional Climate Models: Process Understanding and Stakeholder Information," *Bulletin of the American Meteorological Society*, May 2020.

[5] M. J. Roberts *et al.*, "The benefits of global high-resolution for climate simulation: process understanding and the enabling of stakeholder decisions at the regional scale," *Bulletin of the American Meteorological Society*, vol. 99, no. 11, pp. 2341–2359, 2018.

[6] M. Collins *et al.*, "Emerging signals of climate change from the equator to the poles: New insights into a warming world," *Frontiers in Science*, vol. 2, 2024.

[7] Task Force on Regional Climate Downscaling, WCRP, "CORDEX: A coordinated regional climate downscaling experiment," technical report, World Climate Research Programme, Feb. 2009.

[8] P. Bauer *et al.*, "The quiet revolution of numerical weather prediction," *Nature*, vol. 525, pp. 47–55, Sept. 2015.

[9] E. Moreno-Chamarro *et al.*, "The very-high-resolution configuration of the ec-earth global model for high-resmip," *Geoscientific Model Development*, vol. 18, no. 2, pp. 461–482, 2025.

[10] J. Hoffmann *et al.*, "Destination Earth – A digital twin in support of climate services," *Climate Services*, vol. 30, p. 100394, 2023.

[11] J. Kontkanen *et al.*, "Climate digital twin to support climate change adaptation efforts," in *EGU General Assembly 2023*, (Vienna, Austria), Apr. 2023. EGU23-13018.

[12] Destination Earth, "Destination earth's climate change adaptation digital twin: operationalising multi-decadal climate projections," May 2024. News article; accessed 27 June 2025.

[13] F. Molteni *et al.*, "The ecmwf ensemble prediction sys-

tem: Methodology and validation," *Quarterly Journal of the Royal Meteorological Society*, vol. 122, no. 529, pp. 73–119, 1996.

[14] T. N. Palmer, "The ecmwf ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years," *Quarterly Journal of the Royal Meteorological Society*, vol. 144, no. S1, p. –, 2018.

[15] CASTIEL-2 Consortium, "Access to the supercomputers of the ju – industrial usage." Technical report, EuroCC ACCESS, July 2023.

[16] K. Bi *et al.*, "Pangu-Weather: A 3D High-Resolution Model for Fast and Accurate Global Weather Forecast," Nov. 2022. arXiv:2211.02556 [physics].

[17] M. Reichstein *et al.*, "Graphcast: A graph neural network for global weather forecasting," *Science*, vol. 382, no. 6651, p. eabq8145, 2023.

[18] C. Wong, "How AI is improving climate forecasts," *Nature*, vol. 628, pp. 710–712, Mar. 2024.

[19] Z. B. Bouallègue *et al.*, "The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context," *Bulletin of the American Meteorological Society*, June 2024.

[20] S. Materia *et al.*, "Artificial intelligence for climate prediction of extremes: State of the art, challenges, and future perspectives," *WIREs Climate Change*, vol. n/a, no. n/a, p. e914, 2024.

[21] G. Camps-Valls *et al.*, "Artificial intelligence for modeling and understanding extreme weather and climate events," *Nature Communications*, vol. 16, 2025.

[22] T. Rackow *et al.*, "Robustness of AI-based weather forecasts in a changing climate," Sept. 2024. arXiv:2409.18529 [physics].

[23] M. D. Wilkinson *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, p. 160018, 2016.

[24] D. Manubens-Gil *et al.*, "Seamless management of ensemble climate prediction experiments on HPC platforms," in *2016 International Conference on High Performance Computing & Simulation (HPCS)*, pp. 895–900, July 2016.

[25] N. D. Brenowitz *et al.*, "A Practical Probabilistic Benchmark for AI Weather Models," Nov. 2024. arXiv:2401.15305 [physics].

[26] T. Selz and G. C. Craig, "Can Artificial Intelligence-Based Weather Prediction Models Simulate the Butterfly Effect?," *Geophysical Research Letters*, vol. 50, no. 20, p. e2023GL105747, 2023.

[27] D. Kochkov *et al.*, "Neural general circulation models for weather and climate," *Nature*, vol. 632, pp. 1060–1066, Aug. 2024.

[28] S. Lang *et al.*, "AIFS – ECMWF's data-driven forecasting system," Aug. 2024. arXiv:2406.01465 [physics].

[29] F. Banchelli *et al.*, "Introducing marenostrum5: A european pre-exascale energy-efficient system designed to serve a broad spectrum of scientific workloads," *arXiv preprint arXiv:2503.09917*, Mar. 2025.

[30] S. Lang *et al.*, "AIFS-CRPS: Ensemble forecasting using a model trained with a loss function based on the Continuous Ranked Probability Score," Dec. 2024. arXiv:2412.15832 [physics].

[31] H. Hersbach *et al.*, "ERA5 hourly data on single levels from 1940 to present." `https://doi.org/10.24381/cds.adbb2d47`, 2023. Accessed on 03-07-2025.

[32] D. B. Stephenson and F. J. Doblas-Reyes, "Statistical methods for interpreting monte carlo ensemble forecasts," *Tellus A: Dynamic Meteorology and Oceanography*, vol. 52, no. 3, pp. 300–322, 2000.

[33] T. Rackow *et al.*, "Multi-year simulations at kilometre scale with the Integrated Forecasting System coupled to FESOM2.5 and NEMOv3.4," *Geoscientific Model Development*, vol. 18, pp. 33–69, Jan. 2025.

[34] Deutsches Klimarechenzentrum (DKRZ), "The ssp scenarios — cmip6 climate simulations," n.d. Accessed: 2025-06-24.

[35] M. Nurisso *et al.*, "Destine-climate-dt/aqua: v0.13.7," 2025. Accessed: 2025-06-24.

[36] J. Corbalan *et al.*, "Energy optimization and analysis with ear," in *2020 IEEE International Conference on Cluster Computing (CLUSTER)*, pp. 464–472, 2020.

[37] M. Alexe *et al.*, "Data-driven ensemble forecasting with the aifs," *ECMWF Newsletter*, vol. 181, pp. 32–37, Oct 2024.

[38] https://github.com/ainagaya/auto-UQ4ClimateDT/tree/main

[39] Parameterizations are a method to replace processes that are too small-scale or complex to be physically represented in the model by a simplified process.