Enabling Fundamental Kernels for Ensemble Climate Simulations in the Era of AI Accelerators

L. Breschi ✉, F. Vella
University of Trento, Italy
Correspondence to: lorenzo.breschi@unitn.it

hicrest.unitn.it — www.linkedin.com/company/hicrest-laboratory

Abstract

Despite the necessity of ensemble predictions for weather and climate, current operational Numerical Weather Prediction (NWP) codes, as well as most other Partial Differential Equations (PDE) solvers, compute each ensemble member sequentially, and so it does not scale on modern supercomputers.

Data from nearby spatiotemporal points, as well as data from different nearby ensemble members, is highly correlated, recent works suggest that weather data can be successfully compressed with a minimal loss of information.

We plan to achieve a speedup over the sequential generation of ensemble members by computing all members at the same time and so using the fact that they share common information to compress them and to reduce the number of operations and communications in distributed clusters.

New machine-learning accelerators appear well-suited for this type of parallel computation, and we will design our kernels to be flexible enough for both new and old hardware.

1 Introduction

Partial Differential Equations (PDE) are ubiquitous in physical sciences, especially in computational fluid dynamics (CFD), plasma physics and weather and climate models. Due to the lack on exact mathematical solution for many systems beyond the simple linear models, numerical methods have been an essential for the study of PDEs and the practical scientific and engineering applications [?].

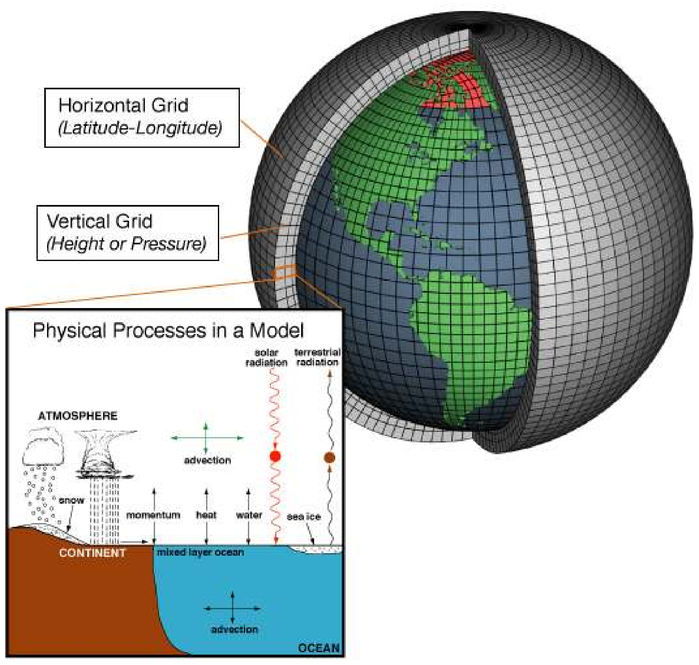
Due to the chaotic nature of the equations of weather, ensemble methods are used in numerical weather prediction (NWP) to provide a probabilistic representation of the future state of the atmosphere [?]. These methods are based on the solution of the same set of PDEs but with slightly different initial conditions and model parametrizations. The ensemble of solutions is then used to provide a probabilistic representation of the future state of the atmosphere, in a Monte Carlo fashion. Due to the limitations of compute power, only a limited number of ensemble members can be run, typically in the order of 10-100 [**?**]. This limits the accuracy of the probabilistic representation of the future state of the atmosphere.

Despite their usefulness, ensemble predictions for PDEs are used only in NWP and each ensemble member is computed sequentially [?]. As far as we know ensemble methods are rarely used in other areas of science where PDEs are used, even in similar fields such as CFD and plasma physics [?].

This work aims to provide a set of fundamental kernels for ensemble predictions of PDEs to show the potentiality of parallel computing in this field.

We plan to achieve a speedup over the sequential generation of ensemble members by computing all members at the same time and using the fact that they share common information to compress data, as suggested by recent works [1, 2], and to reduce the number of operations and communications in distributed clusters.

New ML accelerators seem well suited to this kind of parallel computation [?], and so we will strive to design our kernels with enough flexibility for new and old hardware. While ML seems to provide a potential solution to this problem with a computational speedup, its reliability is still to be proven [?]and so it cannot be used in climate predictions where the reliability is essential. Furthermore, while many ML works claim a far superior speed than the operational methods [3] [?], there is no publicly available speed benchmark to support this claim [?]. Notwithstanding that at the present moment all ML models still need data from NWP models to be trained upon [?].



Atmospheric Model Schematic: PDE solver on a grid plus subgrid parameterizations.

2 State of the Art

The field of numerical weather prediction (NWP) and climate modeling has seen significant advancements in recent years. However, several challenges remain, particularly in the context of ensemble predictions and the efficient use of modern computational resources. Below, we summarize the current state of the art:

- **Compression of Weather Data:** Recent studies have demonstrated the potential of compressing weather and climate data with minimal loss of information. Techniques such as multidimensional compression and hybrid approaches have shown promise in reducing data size while preserving critical features [1, 2].
- **Ensemble Methods:** While advanced ensemble methods exist for data assimilation and ordinary differential equations (ODEs), their application to partial differential equations (PDEs) remains limited. This gap highlights the need for specialized techniques tailored to the unique challenges of PDE-based models.
- **Hardware Compatibility:** Fundamental computational kernels, such as General Matrix Multiply (GEMM) and Fast Fourier Transform (FFT), are widely available and optimized for both legacy and modern hardware. These kernels form the backbone of many scientific computing applications and provide a solid foundation for further development.

3 Research Gap

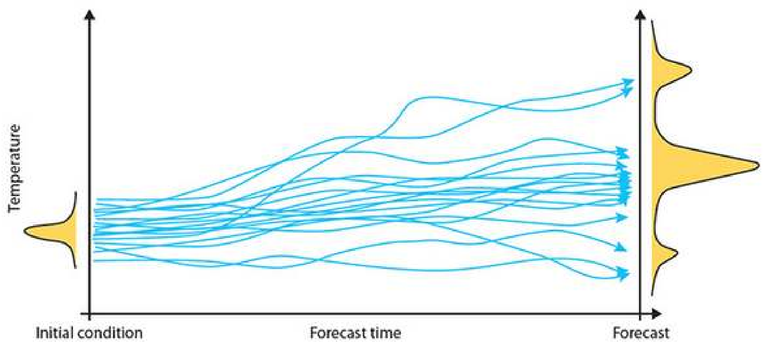
Despite the progress in the field, several critical gaps remain unaddressed. These gaps hinder the full realization of the potential benefits offered by ensemble predictions and modern computational technologies:

- **Sequential Ensemble Predictions:** Current operational NWP codes compute ensemble members sequentially, limiting scalability and efficiency on modern supercomputers.
- **Limited Use of Ensemble Predictions in Other Fields:** While ensemble methods are standard in NWP, their adoption in other fields that rely on PDEs, such as computational fluid dynamics (CFD) and plasma physics, is minimal.
- **Lack of Benchmarking for ML Models:** Machine learning (ML) models claim significant speedups over traditional methods, but there is a lack of publicly available benchmarks to validate these claims. This gap raises questions about the reliability and applicability of ML-based approaches in critical applications.

4 Research Objectives

To address the challenges and gaps identified above, we propose the following research objectives:

- **Development of Fundamental Kernels:** We aim to develop a set of fundamental kernels specifically designed for ensemble predictions of PDEs. These kernels will leverage the inherent parallelism of modern computational architectures to achieve significant speedups.
- **Exploitation of Parallelism:** By treating ensemble members as a single computational entity, we will exploit the parallelism offered by modern clusters and AI accelerators. This approach will reduce redundant computations and improve scalability.
- **Implementation of Compression Techniques:** To further enhance efficiency, we will integrate advanced compression techniques into our workflow. These techniques will reduce data size and computational cost while maintaining accuracy and reliability.
- **Flexibility and Compatibility:** Our kernels will be designed to ensure compatibility with both legacy and modern hardware. This includes optimizing performance for GPUs, TPUs, and other AI accelerators while maintaining compatibility with traditional CPUs.
- **Benchmarking and Validation:** We will conduct rigorous benchmarking and validation to evaluate the performance, reliability, and scalability of our approach. This includes comparisons with existing methods and assessments in various scenarios.



5 Methodology

To address the challenges outlined in the research gap, we propose a comprehensive methodology that combines advanced computational techniques, parallel processing, and data compression strategies. Our approach is structured as follows:

5.1 Parallel Computation of Ensemble Members

The core idea is to compute all ensemble members simultaneously, leveraging the inherent parallelism of modern supercomputers and AI accelerators. By treating the ensemble members as a single computational entity, we can exploit shared information across members to reduce redundant computations. This approach requires the development of specialized kernels that can handle the parallel computation efficiently while maintaining numerical stability and accuracy.

5.2 Data Compression Techniques

Recent studies have demonstrated the potential of compressing weather and climate data with minimal loss of information. We aim to integrate these compression techniques into our workflow to reduce the data size and computational overhead. Specifically, we will explore methods such as:

- Lossless compression algorithms for preserving critical information.
- Lossy compression techniques with controlled error margins to balance accuracy and efficiency.
- Hybrid approaches that combine the strengths of both lossless and lossy methods.

5.3 Kernel Design for Flexibility

To ensure compatibility with both legacy and modern hardware, we will design our kernels to be highly flexible. This involves:

- Implementing hardware-agnostic algorithms that can adapt to different architectures.
- Optimizing performance for AI accelerators, such as GPUs and TPUs, while maintaining compatibility with traditional CPUs.
- Incorporating modular design principles to facilitate future extensions and updates.

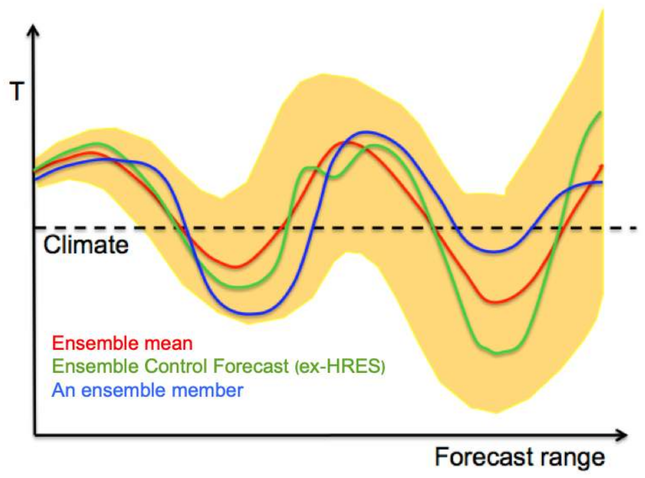
5.4 Benchmarking and Validation

To evaluate the effectiveness of our approach, we will conduct rigorous benchmarking and validation. This includes:

- Comparing the performance of our kernels against existing methods, including ML-based approaches.
- Assessing the reliability and accuracy of our solutions in various scenarios, such as extreme weather events and long-term climate simulations.
- Analyzing the scalability of our methods on distributed systems with varying numbers of nodes.

5.5 Exploration of Adaptive Grids and Asynchronous Methods

As part of our future work, we will investigate the use of adaptive grids in time, space, and ensemble spread to further optimize performance. Additionally, we will explore asynchronous methods for PDE integration to reduce communication delays in distributed systems. These advanced techniques have the potential to significantly enhance the efficiency and scalability of our approach.



6 Further Work

At the moment most predictions are also done on regular grids and each step of the main PDE integration is of the same length, although slow processes are resolved at bigger timesteps [4].

A further line of research that we want to explore in future work is the use of adaptive grids in time, space and ensemble spread to further reduce the computational cost. While adaptive spatial grids are used in many CFD applications [5], adaptive time grids are less common [?]and adaptive ensemble spread grids are not used at all.

Since we need to use a machine with multiple nodes due to the size of the problem, it could be interesting to explore asynchronous methods for the integration of the PDEs, as they could provide a further speedup by reducing communication delays [6, 7, 8]

Methods for integrating Fokker-Planck equation (the equation that describes the evolution of a probability density under an ODE) could also be interesting to explore [9, 10]. While it does not exist a useful equivalent for PDEs [11], we may use it on the discretized version of the PDEs, which is in fact a set of ODEs.

References

[1] Milan Klöwer, Martin Razinger, Juan J. Dominguez, et al. Compressing atmospheric data into its real information content. *Nature Computational Science*, 1:713–724, 2021.

[2] Langwen Huang and Torsten Hoefler. Compressing multidimensional weather and climate data into neural networks, 2023.

[3] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast, 2022.

[4] F. Prill, D. Reinert, D. Rieger, and G. Zängl. Icon tutorial – working with the icon model, 2024.

[5] Anshu Dubey, Ann Almgren, John Bell, Martin Berzins, Steve Brandt, Greg Bryan, Phillip Colella, Daniel Graves, Michael Lijewski, Frank Löffler, Brian O'Shea, Erik Schnetter, Brian Van Straalen, and Klaus Weide. A survey of high level frameworks in block-structured adaptive mesh refinement packages. *Journal of Parallel and Distributed Computing*, 74(12):3217–3227, 2014. Domain-Specific Languages and High-Level Frameworks for High-Performance Computing.

[6] A. Lew and M. Ortiz. *Asynchronous Variational Integrators*, pages 91–110. Springer New York, New York, NY, 2002.

[7] Diego A. Donzis and Konduri Aditya. Asynchronous finite-difference schemes for partial differential equations. *Journal of Computational Physics*, 274:370–392, 2014.

[8] Shubham K. Goswami, Vinod J. Matthew, and Konduri Aditya. Implementation of low-storage runge-kutta time integration schemes in scalable asynchronous partial differential equation solvers. *Journal of Computational Physics*, 477:111922, 2023.

[9] Matthew Dobson, Yao Li, and Jiayu Zhai. An efficient data-driven solver for fokker-planck equations: algorithm and analysis, 2019.

[10] Nan Chen and Andrew J. Majda. Efficient statistically accurate algorithms for the fokker–planck equation in large dimensions. *Journal of Computational Physics*, 354:242–268, February 2018.

[11] Martin Ehrendorfer. The liouville equation and atmospheric predictability. *Predictability of weather and climate*, pages 59–98, 2006.

HICREST

UNIVERSITY OF TRENTO

ICT DAYS