Forecast dispersion and sensitivity in observation-perturbed ensembles of data-driven weather prediction models

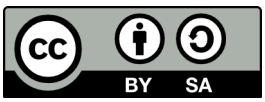
Master's Thesis in
Meteorology and Climate Physics
by

Isabel Pena Sánchez

February 2025

INSTITUTE OF METEOROLOGY AND CLIMATE RESEARCH
KARLSRUHE INSTITUTE OF TECHNOLOGY (KIT)

Supervisor:	Prof. Dr. Peter Knippertz
Co-supervisor:	Prof. Dr. Julian Quinting



Abstract

The application of machine learning (ML) to weather prediction has recently made great advancements, with data-driven models now rivaling traditional numerical weather prediction (NWP) systems. However, most data-driven weather models focus on deterministic forecasting, leaving ensemble forecasting relatively unexplored. One established approach in the Meteorological Service of Canada (MSC) consists in creating an initial condition ensemble by perturbing observations before data assimilation. To explore this further, this study experiments with different approaches of ensemble generation in the TEEMLEAP testbed - a framework developed jointly by the Karlsruhe Institute of Technology (KIT) and the German Weather Service (DWD) on KIT's supercomputer HoreKa - which emulates the operational weather forecasting chain in a simplified yet realistic manner.

In this thesis, uncertainty quantification in data-driven models is investigated by generating probabilistic forecasts from the deterministic data-driven weather prediction model FourCastNet V2. Building on the method of Houtekamer et al. (1996), different ways to systematically perturb the observations before data assimilation with the NWP-based ICON model are tested. The key objectives are: (1) to demonstrate that FourCastNet V2 can be run using ICON-based initial conditions generated within the TEEMLEAP testbed; (2) to compare different initial condition ensemble generation methods by modifying the parameters that control the nature of perturbed observations; and (3) to use the outcome from the most adequate perturbation method to initiate forward integrations with FourCastNet V2, which are then evaluated using probabilistic verification.

This study presents the first successful use of ICON-based initial conditions generated within TEEMLEAP for running a pre-trained version of FourCastNet. Among the tested observation perturbation methods to generate an initial condition ensemble, the approach that changes the random seed in the generation of vertical error profiles proves to be the most effective, demonstrating healthy spread-skill relationships. A probabilistic forecast is then produced by forward integrating the obtained initial condition ensemble using FourCastNet V2. The evaluation reveals an abrupt decrease in ensemble spread during the first 24 hours, leading to a consistent underestimation of uncertainty throughout the entire forecast and indicating that FourCastNet V2 struggles to maintain and grow small-scale structures contained in the ICON-based initial condition ensemble.

Despite the challenges encountered, designing a first data-driven probabilistic forecast prototype with FourCastNet V2 provides an important first step into data-driven ensemble prediction with the TEEMLEAP testbed, which demonstrated its value as a hands-on platform for students to explore ML applications in the weather prediction chain. Future work could include incorporating model error and fine-tuning the perturbed observations. For instance, the MSC inflated initial condition perturbations to achieve a well-calibrated forecast spread. Additionally, forward integration with other data-driven models or alternative ensemble generation methods - such as varying the number of assimilated observations within small ranges - could be explored. Furthermore, investigating scale-dependent error growth in data-driven probabilistic forecasting could provide deeper insights into the representation of uncertainty.

Contents

1 Introduction

Our world is rapidly changing, and societies are facing an increase in the frequency and intensity of high-impact and extreme weather events, as evidenced in daily news headlines (Brunet et al., 2023). Livelihoods can be destroyed, especially those of the poorest communities, underscoring the critical role of early warning systems in effective disaster risk management (Field et al., 2012). This dictates the urgent need for further improvements in weather prediction systems (Brunet et al., 2023).

Over the past three to four decades, global numerical weather prediction (NWP) has seen steady advancements, driven by improvements in the representation of physical processes through higher-resolution models, the optimized data assimilation techniques, more extensive and accurate observations, and enhanced forecast uncertainty estimation using ensemble methods (Bauer et al., 2015). Ensemble prediction is essential in modern NWP, allowing for the quantification of forecast uncertainty (Kalnay, 2002). One relevant method, developed by Houtekamer et al. (1996) and employed by the Meteorological Service of Canada (MSC), involves perturbing the observations before the data assimilation to create an initial condition ensemble, which is then used to generate an ensemble of forecasts. By perturbing the observations, this method accounts for observational uncertainties and ensures that the resulting ensemble of initial conditions represents different possible atmospheric states. In the approach from Houtekamer et al. (1996), observations are perturbed randomly in a manner consistent with observation error statistics, ensuring vertical correlations. While this method has been explored in several studies (Buizza, 2008; Keresturi et al., 2019; Zhang et al., 2010; Goswami et al., 2005), these approaches focus on either random perturbations or varying the number of assimilated observations without perturbing them. This reveals a gap in the exploration of alternative methods for perturbing the observations.

Despite the mentioned achievements in NWP, there is still considerable room for improvement, particularly in optimizing the ensemble systems to enhance forecast reliability, while also addressing the technological challenges related to computational efficiency and data processing capacity (Bauer et al., 2015). Around 2010, three key developments paved the way for a new wave of Artificial Intelligence (AI)-driven research: the advent of graphical processing units (GPUs) for massive parallel processing, the introduction of convolutional neural networks (CNN) for efficient analysis of massive datasets (Krizhevsky et al., 2012), and the availability of large benchmark datasets on the internet (Schultz et al., 2021). These breakthroughs presented an opportunity to drive innovation and improve weather predictions, leading to the first data-driven weather prediction models that emerged in the late 2010s.

The application of machine learning (ML) to weather prediction has recently made great advancements, with data-driven models now rivaling traditional NWP systems in accuracy (ECMWF, 2023). For example, models like FourCastNet (Pathak et al., 2022) or Pangu-Weather (Bi et al., 2022) demonstrated the ability to generate high-resolution global forecasts up to a week ahead, achieving comparable or superior accuracy to NWP models while highly reducing computational costs (Pathak et al., 2022; Bi et al., 2022). However, the exploration of probabilistic data-driven weather forecasts has been relatively limited until recently. Bülte et al. (2024) took a first step toward probabilistic data-driven weather models, by evaluating uncertainty quantification (UQ) methods to generate probabilistic forecasts from the deterministic data-driven model Pangu-Weather. Their work demonstrated the potential of data-driven weather models to generate probabilistic forecasts that are competitive with operational "European Centre for Medium-Range Weather Forecasts (ECMWF)" ensemble.

In 2024, a significant breakthrough was achieved with NeuralGCM, a general circulation model (GCM) capable of generating deterministic and probabilistic weather forecasts, and climate predictions on par with the best ML and physics-based methods (Kochkov et al., 2024). Most recently, the probabilistic data-driven weather model GenCast was introduced, performing probabilistic weather forecasts based on ML that are both more skillful and faster to generate than the leading NWP-based ensemble forecast, ENS of ECMWF (Price et al., 2025).

The rapid advancements in AI-driven atmospheric prediction demands adaptation in academic institutions as well. New teaching curricula, infrastructures, and strategies are essential to prepare the students and researchers for this evolving field. This study aims to contribute to these efforts by using the global "TEstbed for Exploring Machine LEarning in Atmospheric Prediction (TEEMLEAP)" as the foundational framework for all the analyses in this thesis. The TEEMLEAP testbed was developed by an interdisciplinary team of scientists from the Karlsruhe Institute of Technology (KIT) and the German Weather Service (DWD). Implemented on KIT's supercomputer HoreKa, TEEMLEAP mimics the operational weather forecasting chain in a simplified but yet realistic manner, also incorporating interfaces for integrating AI-based components (Wilhelm et al., 2024). One of TEEMLEAP's key objectives is to make advanced weather prediction tools more accessible to students, providing them with a hands-on platform to explore the weather prediction chain, and experiment with ML applications.

Figure 1.1 presents a schematic of this thesis project, adapted from the original TEEMLEAP testbed framework (Wilhelm et al., 2024). As shown in the left panel, the study begins by exploring different ways to systematically perturb the observations before data-assimilation (middle panel) to generate an initial condition ensemble, building upon the method by Houtekamer et al. (1996). Specifically, this includes applying random perturbations to the observations, varying the number of assimilated observations, and modifying their global statistics.

This project is conducted during the period when ensemble data-driven weather prediction is still in its early stages, offering an opportunity to experiment with ensemble approaches. This is done using the initial condition ensemble generated (represented by the assimilation cycle in Figure 1.1) to initiate forward integrations with FourCastNet V2, a state-of-the-art data-driven model developed

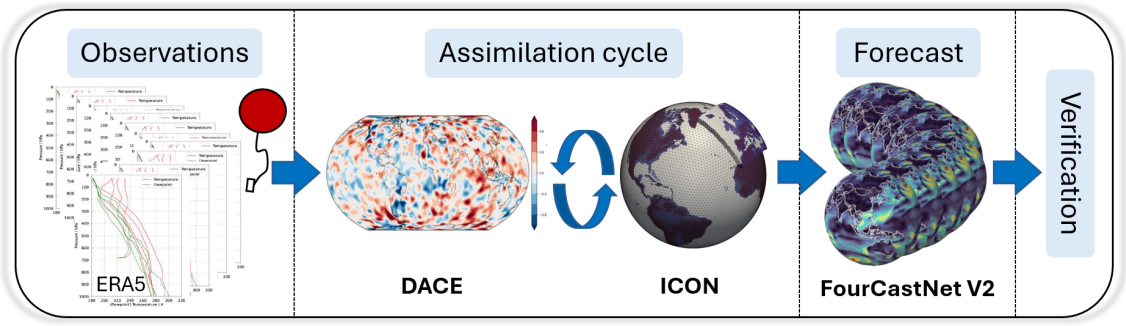


Figure 1.1: Schematic representation of this thesis' approach, adapted from the TEEMLEAP testbed framework (Wilhelm et al., 2024). The workflow begins with an ensemble of perturbed Observations (left), where various methods are explored. These perturbed observations are used in the assimilation cycle to generate ensembles of initial conditions, which are subsequently integrated forward using FourCastNet V2, in the Forecast step. The forecasts are then verified in the final Verification step, which evaluates the ensemble prediction approach.

by Pathak et al. (2022). This marks the first use of FourCastNet V2 for ensemble prediction with NWP-based "ICOsahedral Nonhydrostatic (ICON)" initial conditions from the TEEMLEAP testbed.

Finally, aligning with TEEMLEAP's goal to provide students with a hands-on platform to explore the weather prediction chain and experiment with ML applications, this project represents an important step into data-driven ensemble prediction with the TEEMLEAP testbed: it addresses the technical challenges and designs a first data-driven probabilistic forecast prototype with FourCastNet V2.

Objectives and main goal

The main goal is to develop a sophisticated ensemble generation method for creating probabilistic data-driven forecasts within the TEEMLEAP testbed. This involves exploring different methods for perturbing the observations building on approaches found in the literature, and experimenting with ensemble generation using the deterministic data-driven FourCastNet V2 model. Unlike traditional ensemble prediction systems such as ECMWF's Ensemble Prediction System (EPS) (Molteni et al., 1996), which incorporate both initial condition and model physics perturbations, this study focuses only on perturbations of the initial state. By achieving this goal, we aim to support TEEMLEAP as a valuable learning tool for future students, which can bring data-driven atmospheric prediction closer to educational frameworks. The key objectives of this study are as follows:

1. Prove that the data-driven model FourCastNet V2 can be run with ICON-based initial conditions from the TEEMLEAP testbed.

2. Compare different initial condition ensemble generation methods by modifying the parameters that control the nature of perturbed observations.

3. Select the most adequate observation perturbation method for generating an ICON-based initial condition ensemble with sufficient spread.

4. Use the outcome of the most adequate method, to initiate forward integrations with FourCastNet V2, and evaluate the forecast using probabilistic verification.

The structure of this thesis is as follows: Chapter 2 provides the theoretical background, discussing numerical weather prediction (NWP), with the deterministic and ensemble approaches, as well as data-driven weather prediction models and their application in ensemble prediction. Chapter 3 describes the data and methods employed, including an overview of the ERA5 dataset, the ICON model, FourCastNet V2, and the TEEMLEAP testbed. It also explains the methods for perturbing observations, ensemble generation, and the evaluation metrics used. Chapter 4 presents the results of the experiments, addressing the main research objectives. Section 4.1 demonstrates that ICON initial conditions can be used to run data-driven forecasts. Section 4.2 investigates the different methods for generating ensembles of initial conditions, and Section 4.3 evaluates the performance of the ensemble data-driven forecast. Finally, Chapter 5 summarizes the key findings and the objectives reached, and discusses the potential directions for future research.

2 Background Information

2.1 Numerical Weather Prediction

2.1.1 Deterministic Numerical Weather Prediction

NWP is used to forecast the weather by employing physical models that simulate the Earth's atmosphere by integrating a set of partial differential equations. It has gone through a significant evolution, which is well documented and known as the "quiet revolution" of NWP, and which led to substantial improvements in the accuracy and reliability of forecasts over the past decades (Bauer, 2024).

To construct an effective NWP forecast, several components are necessary: a global observation system, data assimilation, the NWP model, and post-processing and verification stages. These components form a sequential process known as the NWP chain, illustrated in Figure 2.1. The chain begins with meteorological observations, which undergo pre-processing to prepare the data for assimilation. The pre-processed observations and then input into the data assimilation system to generate the initial conditions required for the forecast. This is followed by the prediction step, and the forecasts outputs undergo post-processing and are evaluated during the verification step. Each step will be discussed in more detail in the following.

The foundation of NWP begins with a robust global observation system that is largely based on ground stations and weather balloons, and that is also heavily reliant on satellite data, which remains a challenge in terms of assimilation (Kalnay, 2002). These observational inputs are essential to capture the current state of the atmosphere.

Data assimilation is the process of integrating observational data from various sources with a prior short-term model forecast, known as "first guess" or background, to create the best possible representation of the current atmospheric state. This step produces the optimized initial conditions or "analysis", which are crucial for accurate forecasting (Kalnay, 2002). Data assimilation methods vary in complexity. Traditional three-dimensional approaches, like 3D-Var and optimal interpolation (OI), minimize a cost function that measures the distance between the analysis and both the background and observations (Kalnay, 2002). They are computationally efficient but do not account for temporal evolution. More sophisticated techniques, such as four-dimensional methods (4D-Var), do account for temporal evolution but at a higher computational cost (Kalnay, 2002). Methods such as Kalman filters are also utilized to iteratively refine the initial conditions. A particularly relevant technique is the Physical-Space Assimilation System (PSAS), introduced by Cohn et al.

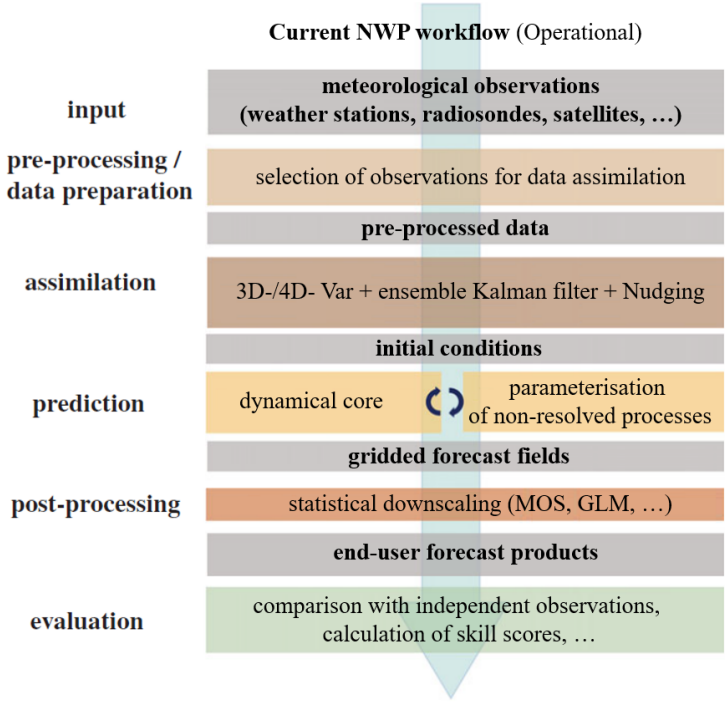


Figure 2.1: Adapted from Schultz et al. (2021) - Schematic representation of the operational NWP workflow, showing the key components of the weather forecasting chain. It begins with meteorological observations as input, which are pre-processed and selected for the data-assimilation step, to generate the initial conditions. The prediction step follows, where the NWP model uses a dynamical core and parameterizes the unresolved processes. The forecast outputs are then post-processed (e.g., statistical downscaling) to produce end-user forecast products. The workflow concludes with the evaluation-verification step.

(1998). PSAS shares similarities with 3D-Var and OI but differs in its approach by performing the minimization of the cost function in the physical space of the observations, rather than in the model space as in the 3D-Var scheme (Kalnay, 2002). The cost function is formulated as:

$$J(\mathbf{w}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}})\mathbf{w} - \mathbf{w}^{\mathrm{T}}\left[\mathbf{y}_{\mathrm{o}} - H(\mathbf{x}_{\mathrm{b}})\right] \tag{2.1}$$

$\mathbf{R}$ is the observation covariance matrix, $\mathbf{H}$ denotes the observation operator, $\mathbf{B}$ is the background error covariance matrix, $\mathbf{y_o}$ is the observation field (a vector of length equal to the number of observations), and $\mathbf{x_b}$ is the model background field. $H(\mathbf{x}_{\mathrm{b}})$ transforms the background field into the observation space using the observation operator (Kalnay, 2002). $\mathbf{w}$ is the intermediate solution vector, given by:

$$\mathbf{w} = \mathbf{R}^{-1}(\delta\mathbf{y}_{\mathrm{o}} - \mathbf{H}\delta\mathbf{x}_{\mathrm{a}}) = \mathbf{R}^{-1}\left[\mathbf{y}_{\mathrm{o}} - H(\mathbf{x}_{\mathrm{a}})\right] \tag{2.2}$$

which accounts for the misfit of the observations $\mathbf{y}_{\mathrm{o}}$ to the analysis $\mathbf{x}_{\mathrm{a}}$, weighted by the inverse of the observation covariance matrix $\mathbf{R}$ (Kalnay, 2002). As will be seen later, the PSAS method is employed in this work during the assimilation phase.

The NWP model employs a dynamical core that is responsible for solving the fundamental physical equations governing atmospheric behavior. These include Partial Differential Equations (PDEs) such as the Navier–Stokes equations for fluid motion, the mass continuity equation (including the effect of the Earth's rotation), the first law of thermodynamics, and the ideal gas law (Warner, 2011). They represent the full set of prognostic equations upon which the change in space and time of wind, pressure, density and temperature is described in the atmosphere (Warner, 2011). Because of the mathematical impossibility of obtaining analytical solutions, NWP models use numerical methods to discretize in space and time, approximating these equations on a grid. This discretization introduces the distinction between resolved scales (large-scale atmospheric movements captured by the grid) and unresolved scales of motion (those that occur at scales too small to be explicitly resolved by the model grid) (Kalnay, 2002).

In NWP, processes that cannot be explicitly resolved by the model are referred to as "subgrid-scale processes". These include turbulent motions with scales ranging from a few centimeters to the size of the model grid, as well as processes that occur at a molecular scale, such as condensation, evaporation, friction and radiation. To represent these subgrid-scale processes, parameterizations are employed, formulating their collective effects in terms of their interaction with the resolvable-scale variables (Warner, 2011). Parameterizations play a fundamental role in determining predictive skill of NWP models, because they determine key aspects of the simulated weather (Warner, 2011).

Verification and post-processing are the final elements in the process chain of modern NWP. Forecast verification is the process of assessing the accuracy and skill of NWP model outputs by comparing them against observed data using metrics like mean error, root mean square error, and correlation coefficients (Wilks, 2011). This step helps to identify model biases and guides improvements in its performance. The post-processing is done to correct systematic errors, estimate uncertainties, derive additional quantities, or for performing spatio-temporal interpolation (Wilks, 2011).

Since 1950s, advancements in high-performance computing (HPC) have enabled the development of more complex and computationally demanding models (Bauer, 2024). Furthermore, as discussed in the following section, to quantify forecast uncertainty and provide probabilistic predictions, current NWP models are typically run multiple times from slightly perturbed initial conditions and perturbed models physics, producing an ensemble of predictions.

2.1.2 Ensemble Numerical Weather Prediction

Early in the 20th century, Poincaré recognized that forecasts of nonlinear systems could vastly differ if small perturbations were applied to the initial conditions, suggesting a fundamental limit to predictive skill (Poincaré, 1914). Lorenz formulated this understanding in a series of remarkable papers (Lorenz, 1963, 1965, 1969), founding chaos theory and demonstrating that even with perfect models and observations, the chaotic nature of the atmosphere imposes a finite predictability limit of about two weeks. This insight emphasized the need to account for the growth of initial condition uncertainties, their evolution, and model imperfections (Palmer, 2006).

Overview of the Lorenz attractor

The Lorenz attractor serves as a conceptual framework for understanding atmospheric predictability (Lorenz, 1963). Its two lobes (which can be seen in any panel of Figure 2.2) represent two distinct weather regimes, such as westerly zonal flow and blocked flow in mid-latitudes, with transitions between the lobes corresponding to shifts in weather regimes (Lynch, 2006). Palmer (1993) used the Lorenz attractor to study the predictability in ensembles of initial states centered around various points within one lobe of the attractor. Over around 5 days of integration (0.5 nondimensional time units), the predictability was shown to vary substantially depending on the initial positions within the attractor. Figure 2.2 highlights three cases of predictability:

- High predictability (a): initial conditions on the left-hand lobe of the attractor evolve together to the right, implying little impact of initial error.
- Moderate predictability (b): the ensemble approaches a region where adjacent trajectories split, leading to increased dispersion and reduced predictability.
- Low predictability (c): initial conditions close to the splitting region diverge rapidly, resulting in an immediate loss of predictive skill.

This analysis underscores the chaotic nature of the atmosphere, where small initial errors grow rapidly, limiting deterministic predictability (Palmer, 1993). The sensitivity to small initial errors is also flow-dependent, which means it is crucial to carefully select initial conditions and to know the reliability of a forecast at any given time (Lynch, 2006).

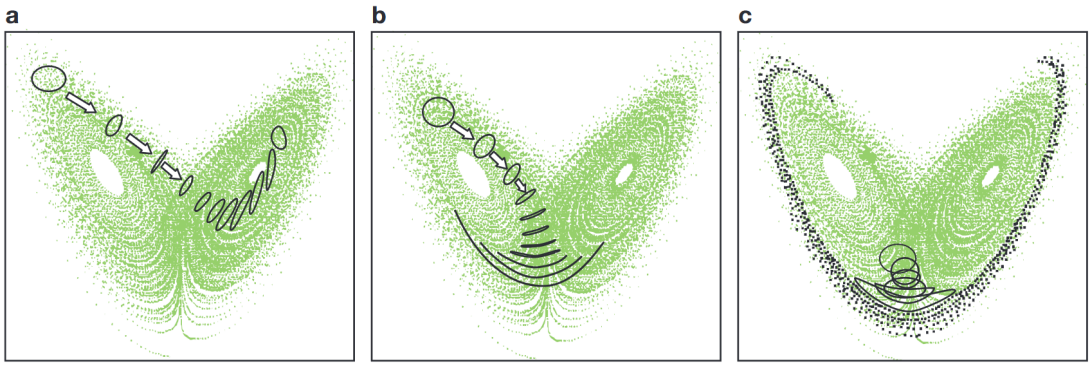


Figure 2.2: From Palmer et al. (2006) - Evolution of an ensemble of initial points (black) within the Lorenz attractor (green), for three sets of initial conditions. The ensemble is plotted every 12 h and integrated for 5 days in equivalent real time.

Ensemble NWP

Building on the work from Lorenz (Lorenz, 1963), NWP has evolved from a *deterministic* approach, based on single numerical integrations, to a *probabilistic* one, that uses ensembles of numerical integrations to estimate the probability distribution of forecast states (Palmer, 2006). This shift enables today's NWP to quantitatively assess the degree of confidence users should have in any particular forecast (Palmer, 2006).

Ensemble forecasting addresses the inherent uncertainty and chaotic nature of the atmosphere by simulating multiple forecasts with different initial states and/or model formulations (Palmer, 2019a). By capturing a range of possible outcomes, ensemble methods provide a probabilistic view of future atmospheric states, rather than a single, deterministic forecast. This is a nontrivial task because the actual uncertainty is flow-dependent and thus varies from day to day (Leutbecher and Palmer, 2008). Fundamental sources of uncertainty in NWP forecasts include initial condition errors, model errors, and the inherent chaotic behavior of the atmosphere. Both initial condition and model uncertainties impact the forecast, and they are interdependent; initial condition error is inseparable from model error for a real physical system like the atmosphere (Leutbecher and Palmer, 2008).

In practice, ensemble members are generated by introducing perturbations to the initial conditions and the model physics, representing sources of error in the analysis and the model (Warner, 2011). Current NWP models, however, operate within vast, high-dimensional phase spaces $N \approx 10^7$, making it impossible to perturb all degrees of freedom individually (Leutbecher and Palmer, 2008). Therefore, ensemble methods are designed to create representative ensemble members that characterize the outer parts of uncertainty space, capturing the variety of possible future atmospheric states. Rather than relying on statistical approximations, ensemble forecasting applies many nonlinear, physical realizations of the atmospheric system, in which observational information refines the forecasts, reducing uncertainty (Palmer, 2006). For effective initial conditions perturbations in NWP ensemble prediction, several characteristics are desirable. First, perturbations should be realistic in the sense that they represent possible atmospheric perturbations consistent with uncertainties in the observed properties of the atmosphere. Second, they should capture the fastest growing modes out of these, which are both realistic and compatible with uncertainty (Kalnay, 2002).

Figure 2.3 presents a schematic of model-state trajectories (dashed lines) for an 8-member ensemble forecast simulation, initialized from perturbed initial conditions in a two-dimensional phase-space. Each dimension of the phase space corresponds to a dependent variable of the system, and the trajectories represent the temporal evolution of the system's states. The initial atmospheric states are defined by different phase-space coordinates, marked by open circles. These circles also mark the states' projection at an intermediate forecast step, and at the final forecast. At each stage, the ensemble mean is denoted by "x". Initially, the error growth appears approximately linear. However, as the forecast progresses, two trajectories diverge significantly until the final forecast projection. This behavior aligns with the discussion in Figure 2.2, where most ensemble forecasts predict one of the possible weather regimes - remaining within one of the phase-space lobes - while two members transition to a different regime, shifting to the opposite lobe. Furthermore, the forecast initiated from the ensemble mean (solid arrow) differs from the time-dependent behavior of the ensemble mean itself. This shows that an atmospheric model is a highly nonlinear function (Wilks, 2011), and the best forecast does not result from the best estimate of the initial conditions (the ensemble mean).

Ensemble prediction systems (EPSs) have recently become well-established at major NWP centers, each employing distinct configurations for the initial-state perturbations. Below, the three most

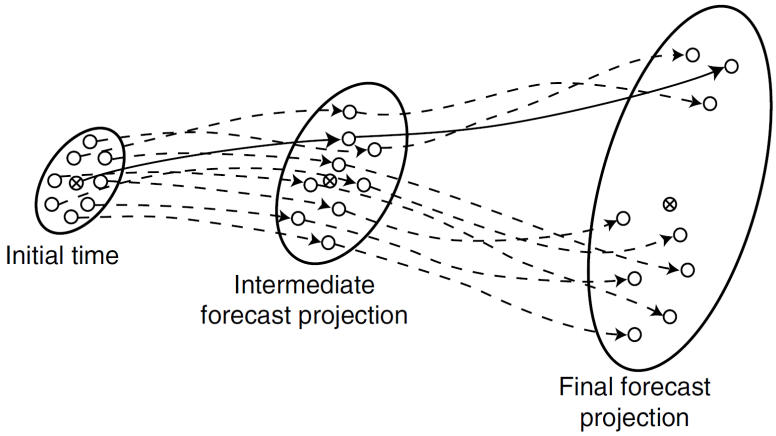


Figure 2.3: From Warner (2011) - Schematic of model-state trajectories for simulations from an 8-member ensemble initialized from perturbed initial conditions.

important EPSs are introduced, including a comparison done by Buizza et al. (2005) of their performance by using 10-member ensembles over a 10-day forecast period.

- The ECMWF EPS uses singular vectors (SV) to represent the leading linear perturbation dynamics over a short-range forecast trajectory. These SV identify the directions in phase space associated with maximum perturbation growth during the early parts of the forecast period (Molteni et al., 1996). Buizza et al. (2005) found that the ECMWF EPS exhibited the highest overall forecast skill, though this was primarily due to its higher model resolution and better data assimilation system, rather than the SV method itself.

- The NCEP (National Center for Environmental Prediction) ensemble is based on the breeding method. Here, a difference field between two nonlinear forecasts is carried forward and periodically scaled down in relation to atmospheric analysis fields (Toth and Kalnay, 1997). In the comparison from Buizza et al. (2005), the NCEP EPS was competitive during the first few days of forecast, best representing small-scale error patterns.

- The EPS of the Canadian Meteorological Centre (CMC), which operates under the MSC, creates initial perturbations through the “ensemble Kalman filter (EnKF)”, in which assimilated observations are perturbed by pseudo-random noise that represents observational error (Houtekamer and Zhang, 2016). It also accounts for model error usign stochastic parameterization techniques (Houtekamer and Zhang, 2016). According to Buizza et al. (2005), the MSC EPS showed superior statistical reliability at longer leadtimes, benefiting from the use of multiple model versions to represent model uncertainty.

The study from Buizza et al. (2005) highlights that the performance of an EPS strongly depends on the quality of the data assimilation system and the numerical model used to generate forecasts. In this thesis, a method similar to the CMC approach is applied, by perturbing observations with vertically correlated noise to generate ensemble members (Houtekamer et al., 1996). This approach will be detailed in Section 3.2.2.

To account for model error in ensemble prediction, it is important to represent the uncertainties from parametrization processes. Without representing these uncertainties, the ensemble would be under-dispersive, meaning that the spread of the ensemble members is too small to represent the actual forecast uncertainty (Leutbecher and Palmer, 2008). This would require other uncertainties, such as those from observational errors, to be inflated to prevent under-dispersion (Leutbecher and Palmer, 2008). Additionally, it was already mentioned that initial error and model error are interdependent, since the NWP model itself is used to assimilate observations.

There are currently three general methods for representing model error in ensemble systems: the multi-model ensemble, the perturbed parameter ensemble, and the stochastic-dynamic parameterization (Leutbecher and Palmer, 2008). A multi-model ensemble involves using a set of quite different parameterization schemes and numerical methods developed by different institutes. In the perturbed parameter ensemble, a single model is run with variations in a certain parameter within the parameterization scheme. The stochastic-dynamic parametrization approach introduces stochastic elements to represent the subgrid-scale variability, without assuming that any single parameterization can provide a deterministic solution (Leutbecher and Palmer, 2008).

2.2 Data-driven Weather Prediction

Scientific and technological developments in NWP have significantly improved weather forecast skill over the past 40 years (Bauer et al., 2015). However, we are now reaching the upper limits of processing speed for the same cost, as well as the limits of the computing performance of complex NWP codes on the current technologies (Bauer, 2024). Further enhancements ultimately need larger and more costly computing infrastructures, which are likely to become unaffordable for operational weather centers. In light of these challenges, ML presents promising alternatives. Driven by data rather than traditional physical modeling, it has the potential to open new avenues for efficient weather forecasting (Bauer, 2024). The data-driven approach is discussed in this section.

2.2.1 Data-driven Models for Weather Prediction

As mentioned in Section 2.1.1, conventional NWP methods model atmospheric physics using PDEs and solve them with numerical simulations, incorporating parameterizations to represent sub-grid scale processes (Kalnay, 2002). Although being very complex, these PDE-based models are often limited by significant computational costs. Additionally, quantifying forecast uncertainty through an ensemble of predictions requires further computational resources, reducing the achievable spatial and temporal resolution, and limiting the ensemble size (Bülte et al., 2024).

To alleviate the above burden, recent advancements in ML-based weather prediction (MLWP) offer a promising alternative to traditional NWP. Studies have shown that MLWP can provide more accurate and efficient non-probabilistic forecasts (Price et al., 2024). This rise of data-driven methods in weather forecasting has been fueled by the availability of massive datasets, advances in

computational hardware, and open-source deep-learning (DL) software frameworks (Düben et al., 2021). In this work, the MLWP approaches are referred to as data-driven models.

ML has a wide range of potential application areas throughout the workflow of NWP and is widely explored within ECMWF initiatives (Düben et al., 2021). While current research focuses on the forecasting part, ML can be applied in other steps of the prediction chain, such as anomaly detection in observations, bias correction in data assimilation, and feature detection during post-processing (Düben et al., 2021). Figure 2.4 illustrates three variations of the NWP workflow. The first (left) represents the traditional NWP, as previously discussed in Section 2.1.1 and shown in Figure 2.1. The center shows a hybrid workflow where specific parts of the NWP prediction chain are substituted with ML/Deep-Learning (DL) approaches. The third (right) shows an end-to-end DL workflow, where all core parts are replaced by a single Deep Neural Network (DNN), which maps meteorological observations directly to end-user forecast products, as investigated in studies such as Grover et al. (2015). In this study, however, we focus on exploring a hybrid workflow: performing data assimilation using 3D-Var-based PSAS and using a DNN for the prediction phase.

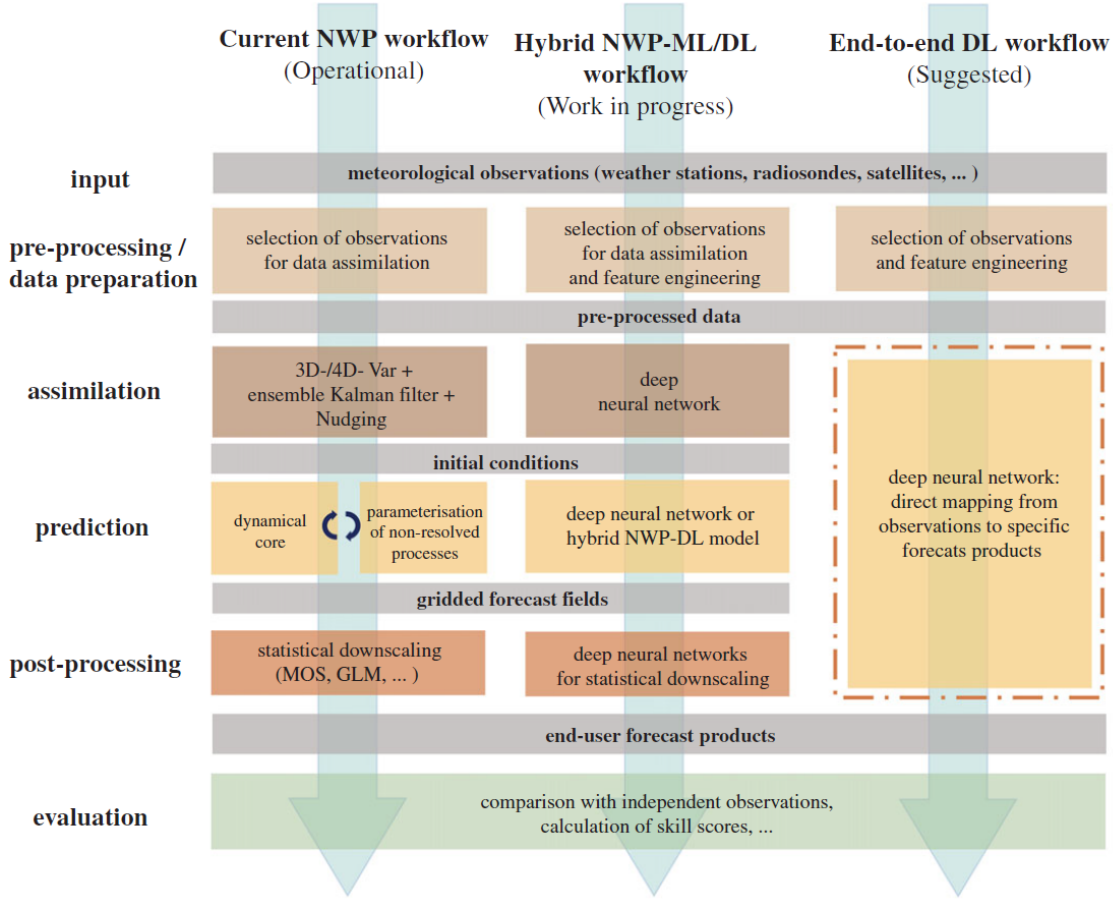


Figure 2.4: From Schultz et al. (2021) - Idealized workflows of current NWP (left), next-generation weather prediction with individual components substituted or augmented by ML and DL techniques (centre), and a purely data-driven DL weather forecasting system (right).

Data-driven models offer a fundamentally different approach by learning the algorithms and statistical patterns from large volumes of past data. The methodology involves training a DNN to model the relationship between the input (ECMWF ERA5 reanalysis data, detailed later in Section

3.1.1, at a given point in time) and the output (reanalysis weather data at the target point in time) (Pathak et al., 2022). Through multiple training layers and learnable parameters, data-driven models capture the statistical dependencies directly from abundant training data, eliminating the need for physical parameterizations and their associated biases (Rasp et al., 2018). After the training, they generate forecasts starting from initial conditions provided by a pre-existing NWP model. Thus, as shown in Figure 2.4, the data assimilation step is still necessary, but it is performed externally by the physical NWP model used to generate the initial conditions.

Fundamental advances in recent years have established purely data-driven models as viable alternatives to traditional NWP, often convincingly outperforming state-of-the-art NWP deterministic systems (Bülte et al., 2024). Key deterministic models include:

- FourCastNet (Pathak et al., 2022), which employs a modified vision transformer (Guibas et al., 2022), for high resolution (0.25º) predictions at 6-hour intervals. It is used in this work, concretely the latest version FourCastNet V2 (Bonev et al., 2023), as discussed further in Section 3.1.3.

- Pangu-Weather (Bi et al., 2022), a model based on vision transformers, which achieves a deterministic forecast accuracy that outperforms the High-RESolution configuration of the Integrated Forecasting System (HRES IFS) from the ECMWF.

- GraphCast (Lam et al., 2023), a Graph Neural Network (GNN)-based model which attains 0.25° horizontal resolution, has a forecast skill and efficiency compared to HRES, and performs well on severe event forecasting.

- NeuralGCM (Kochkov et al., 2024), unlike the previously mentioned models, is a hybrid data-driven General Circulation Model (GCM). It combines a traditional differentiable dynamical core for solving the discretized governing dynamical equations, with a neural network-based physics module that parameterizes physical processes. It generates forecasts of deterministic weather, including also an ensemble version, on par with the best ML and physics-based methods.

To better understand the resources required to train a data-driven model, we can consider FourCastNet as an example. The model was trained on a cluster of 64 GPUs, each with 80 GB of memory. The total training time was approximately 16 hours, resulting in 1,024 GPU-hours (64 GPUs × 16 hours). Figure 2.5 illustrates the capability of FourCastNet with an example forecast of an extreme event visualized using NVIDIA's FourCastNet interactive tool, available from (NVIDIA, 2023). The forecast of wind speed shows the extreme event of Typhoon Hagibis, which struck Japan in October 2019. This powerful typhoon, with wind speeds reaching up to 195 $\mathrm{km\,h^{-1}}$, caused extensive damage estimated at 17.3 billion USD.

Recent studies investigate the capability of data-driven models to provide reliable forecasts of extreme events. In Pasche et al. (2025), GraphCast, PanguWeather, and FourCastNet were compared to ECMWF's HRES in three case studies: the 2021 Pacific Northwest heatwave, the 2023 South Asian humid heatwave, and the North American winter storm in 2021. They found that while ML

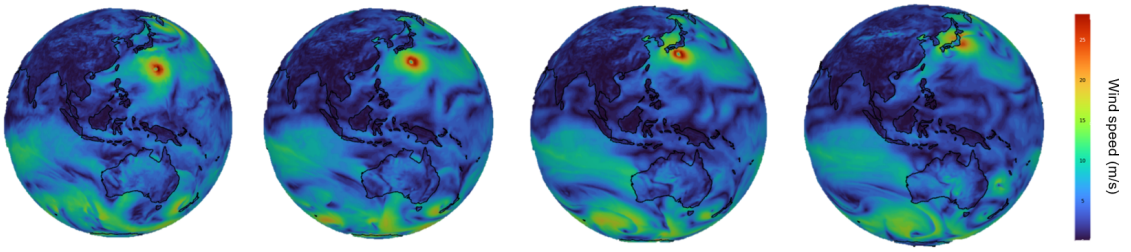


Figure 2.5: Visualization of a FourCastNet forecast from NVIDIA's tool (NVIDIA, 2023), for Typhoon Hagibis, which struck Japan in October 2019, reaching wind speeds of up to 195 km h^{-1}, and causing significant damage. The forecast, generated using NVIDIA's interactive tool, depicts the evolution of wind speed (m s^{-1}) during the event.

models can match HRES in some cases, they struggle with extrapolating extreme conditions, as seen in the Pacific Northwest heatwave, and tend to underestimate the highest danger levels in humidity-driven extremes. However, for the winter storm, PanguWeather and GraphCast outperformed HRES at several lead times, highlighting both the potential and limitations of data-driven models (Pasche et al., 2025).

2.2.2 Ensemble Prediction with Data-driven Models

For a chaotic atmosphere with uncertain initial conditions, ensemble weather forecasting is essential to quantify the likelihood of extreme events and improve the reliability and accuracy of long-term predictions. However, most data-driven forecasting efforts have focused on deterministic forecasts only (Bülte et al., 2024). This approach produces unrealistically blurry predictions when optimized for multi-day forecasts (Kochkov et al., 2024), making it impossible to quantify forecast uncertainties, which is crucial for optimal decision making (Gneiting and Katzfuss, 2014). As explored in this section, data-driven models can enable the generation of large ensembles at low computational cost, which is very valuable for probabilistic forecasting and data assimilation (Pathak et al., 2022). To illustrate this, Table 2.1 provides a comparison of the computational costs for running a 100-member 24-hour ensemble forecast at 18 km resolution using the data-driven FourCastNet versus the traditional IFS model.

Table 2.1: Comparison of computational costs for a 100-member 24-hour ensemble forecast at 18 km resolution using FourCastNet (FCN) and IFS, demonstrating the efficiency of FourCastNet (Pathak et al., 2022).

24-hour 100-member ensemble forecast	FCN - 18 km	IFS
Nodes required	2	3,060
Node-seconds	22	984,000
Energy consumed (kJ)	22	271,000

There are several approaches to generate probabilistic predictions from deterministic data-driven weather models (Bülte et al., 2024). We consider two main approaches for UQ in data-driven weather models. A schematic overview is provided in Figure 2.6.

Post-hoc (PH) UQ approaches. Represented in Figure 2.6a, these methods only require a training dataset of deterministic forecasts and corresponding observations. PH approaches operate directly on a given deterministic forecast from a data-driven model. They use statistical or ML techniques to learn from past pairs of forecasts and observations how to best generate a probabilistic forecast from the deterministic input (Bülte et al., 2024). From a meteorological perspective, this can be viewed as a post-processing task (Vannitsem et al., 2021).

Initial condition (IC)-based approaches. Represented in Figure 2.6b, these approaches generate ensemble forecasts by running a data-driven model multiple times based on a number of slightly different initial conditions. The techniques for generating initial condition ensembles include:

- **Gaussian noise perturbations (GNP)**: As shown in Figure 2.6b (left panel of the UQ methods), Pathak et al. (2022) proposed perturbing the initial weather state from the ERA5 dataset with Gaussian random noise to generate an ensemble of $E = 100$ perturbed initial conditions for FourCastNet. Given an initial condition $X_{\text{true}}(k)$, they generate an ensemble of E perturbed initial conditions $\left\{X^{(e)}(k) = X_{\text{true}}(k) + \sigma\varepsilon\right\}_{e=1}^{E}$ by adding a normally distributed random variable $\varepsilon \sim \mathcal{N}(0,1)$ of the same shape as the initial conditions data, with unit mean and variance. This results in a set of initial conditions used for ensemble data-driven forecasts (Pathak et al., 2022).

- **Initial conditions from the ECMWF ensemble model**: It is based on the work of Buizza (2008), which details the generation of ensembles of initial conditions within the ECMWF prediction system for use in ensemble forecasts (Figure 2.6b, middle panel of the UQ methods). These initial conditions are produced using two different methods: SV (discussed in Section 2.1.2) and Ensemble Data-Assimilation (EDA) Perturbations.

  The EDA approach involves perturbing the observations, the atmospheric boundary conditions such as sea-surface temperature, and the model physics during the assimilation cycles. Buizza (2008) explores a combined approach that integrates both SV and EDA-based perturbations. Their study proposes using the resulting set of perturbed initial conditions, generated by the physics-based ECMWF model, as inputs for ensemble forecasts (Buizza, 2008). Since 2010, ECMWF has operationally used this combined SV-EDA approach to generate initial condition ensembles by running multiple 4D-Var data assimilations (Lang et al., 2019).

- **Random field perturbations (RFP)**: Magnusson et al. (2009) suggested generating initial conditions based on perturbations derived from randomly selected past atmospheric states (Figure 2.6b, right panel of the UQ methods). This method constructs perturbations as the difference between two independently analyzed atmospheric states, rescaled to an appropriate amplitude for ensemble perturbations (Magnusson et al.,

2009). The perturbations are calculated as follows, where $\mathbf{x}_{\mathbf{m}}^{\mathbf{a}}$ represents the difference between the model state vector of an ensemble member m and the analysis initially a:

$$\mathbf{x}_{\mathbf{m}}^{\mathbf{a}} = \alpha \frac{(\mathbf{a}_{\mathbf{d}_1} - \mathbf{a}_{\mathbf{d}_2})}{|\mathbf{a}_{\mathbf{d}_1} - \mathbf{a}_{\mathbf{d}_2}|_{E_{\text{tot}}}}$$

where α is a tuning constant to achieve sufficient ensemble dispersion, and $\mathbf{a}_{\mathbf{d}_1}$, $\mathbf{a}_{\mathbf{d}_2}$ are the state vectors of the analysis from the dates d1 and d2, respectively. The dates are chosen from the same season and different years. The resulting difference field is normalized so that all perturbations have the same amplitude (Magnusson et al., 2009).

- **Observation perturbation method**: In this study, the observation perturbation method introduced by Houtekamer et al. (1996) is employed to generate an initial condition ensemble, with further details provided in Section 3.2.2. The generation of the initial condition ensemble involves parallel data assimilation cycles using the ICON model (detailed in Section 3.1.2) from DWD. These cycles are performed with systematically perturbed observations, designed to maintain vertical correlations and adhere to global statistical properties (Houtekamer et al., 1996).

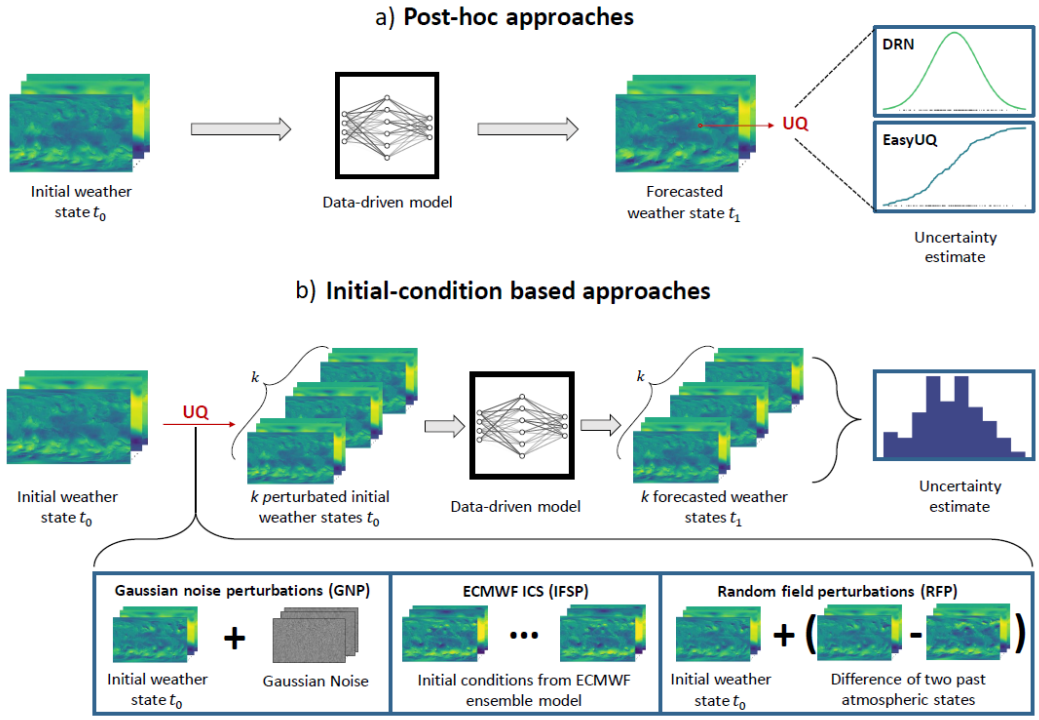


Figure 2.6: Schematic illustration from Bülte et al. (2024) showing the different UQ approaches to generate probabilistic forecasts from deterministic data-driven weather models. The figure distinguishes between PH UQ approaches and IC-based approaches. For the IC-based approaches, the bottom panels illustrate three methods for UQ: GNP on the left, Initial Conditions from the ECMWF ensemble model in the middle, and RFP on the right.

Weyn et al. (2021) demonstrated that large data-driven ensembles improve subseasonal-to-seasonal (S2S) forecasts compared to operational NWP models, which typically incorporate only a small

number of ensemble members (Pathak et al., 2022). Large ensembles also enhance the accuracy of data-driven predictions for extreme weather events in both short- and long-term forecasts (Chattopadhyay et al., 2020).

Recently, the GenCast data-driven diffusion model (Price et al., 2025) has emerged as a new approach for probabilistic forecasting. A diffusion model transforms noisy data into clean data while directly learning to generate probability distributions (Ho et al., 2020), eliminating the need for UQ methods explained above. GenCast produces 15-day ensembles of stochastic global forecasts at a 0.25º resolution. Its probabilistic forecasts are well-calibrated and outperform the ECMWF ensemble forecast (ENS), offering as well enhanced accuracy in predicting extreme weather events and tracking tropical cyclones (Price et al., 2025). Furthermore, Neural GCM (Kochkov et al., 2024), already introduced in Section 2.2.1, produces ensemble forecasts that achieve competitive performance with purely data-driven models for one- to ten-day forecasts, and with the ECMWF ensemble prediction for one- to fifteen-day forecasts (Kochkov et al., 2024). To generate the ensemble, NeuralGCM introduces stochasticity by incorporating additional random fields into the inputs of its neural network components, allowing the model to produce a range of possible weather scenarios (Kochkov et al., 2024).

Until now, NWP methods have led in forecast accuracy and resolution, while data-driven models showed their advantages in efficiency. However, these recent developments open a new chapter in operational weather forecasting, supporting the ability of AI-based methods to accurately capture high-dimensional and complex distributions.

3 Data and Methods

This chapter provides an overview of the data and methods used for the study. It begins with a detailed description of the datasets and models in Section 3.1, including the ERA5 reanalysis dataset, the ICON model for data assimilation, and the FourCastNet V2 model for data-driven forecasting. The TEEMLEAP testbed is then explained in Section 3.2: the observation perturbation method to create initial condition ensembles in Section 3.2.2, and the description of the assimilation cycling environment in Section 3.2.3. The methodology behind the experiments conducted as an *extension* of the TEEMLEAP testbed are presented in Section 3.3: an overview of the modified testbed in section 3.3.1, the alternative methods for the generation of initial condition ensembles in Section 3.3.2, and the probabilistic forecast with FourCastNet V2 in Section 3.1.3. Finally, the methods for the evaluation of ensemble reliability are presented in Section 3.4.

3.1 Data Sources and Models

3.1.1 ERA5 Reanalysis Dataset

The dataset used as observations is ERA5, the fifth-generation reanalysis produced by the ECMWF (Soci et al., 2024). ERA5 provides a detailed representation of global climate and weather over the past decades, generated through data assimilation using the ECMWF IFS and observations from the past. The data is accessible online (Hersbach et al., 2023). It covers the globe with a TL639 (31 km) grid and using 137 hybrid sigma/pressure model levels from the surface up to a height of 80 km. The dataset consists of a wide range of atmospheric, ocean-wave and land-surface variables, available hourly from 1940 to present (Soci et al., 2024). For the purpose of this study, the ERA5 data are remapped onto a $0.25^\circ \times 0.25^\circ$ regular latitude-longitude grid for the period from 27 August to 07 October 2022 (Wilhelm et al., 2024). The chosen study period aligns with the cycling period used in the first version of the TEEMLEAP testbed, presented in Wilhelm et al. (2024). This timeframe, transitioning from summer to autumn, is chosen to minimize the effects of missing surface analyses (as sea-surface temperature and snow cover vary only little), which are not included in this version of the testbed (Wilhelm et al., 2024). In this study, the vertical profiles of four meteorological variables - temperature, relative humidity, geopotential height and horizontal wind - are taken from ERA5 reanalysis data to generate pseudo-observation (PSO) profiles for the assimilation, as will be explained in Section 3.2.2. In the testbed, profiles are created on 98 hybrid model levels, from 40–137 (from surface to approximately 24 km altitude). Additionally, the ERA5 climatology for the period 1990–2020 is used for the evaluation of anomalies in predicted data.

3.1.2 The ICON Model

ICON, developed jointly by DWD and the Max Planck Institute for Meteorology, serves as the operational NWP model of DWD (Zängl et al., 2015). The prognostic variables include the horizontal and vertical wind components, density, and virtual potential temperature. The ICON dynamical core is formulated on an icosahedral-triangular Arakawa C grid. Time integration is performed with a two-time-level predictor-corrector scheme that is fully explicit, except for the terms describing vertical sound-wave propagation (Zängl et al., 2015). DWD uses the ICON model on a global scale with a horizontal resolution of 13 km. Over the European region, it uses a refined nest with a resolution of 6.5 km, and around 2.0 km in the local version over Central Europe. In this study, ICON is employed for the assimilation process with PSAS by generating short-term forecasts, serving as the background fields for the assimilation cycles. For this purpose, operational ICON background fields from the DWD database are interpolated to a 26 km horizontal resolution (R03B06) on 90 vertical levels (Wilhelm et al., 2024).

3.1.3 The FourCastNet V2 Model

FourCastNet, short for Fourier ForeCasting Neural Network, is a global data-driven weather prediction model that provides accurate short to medium-range global predictions with a spatial resolution of 0.25º (Pathak et al., 2022). To produce the high-resolution forecasts, it employs a neural network architecture, the Adaptive Fourier Neural Operator (AFNO) model (Guibas et al., 2022). It is particularly suitable for high-resolution inputs and integrates key recent advances in DL into one model (Pathak et al., 2022). The AFNO model employs the Fourier Neural Operator (FNO) learning method, developed by Zhang (2022) to model complex PDE systems, such those governing fluid dynamics (Zhang, 2022). The FNO learns in a resolution-invariant manner, which allows it to generalize well across different resolutions (Kabri et al., 2023). Additionally, AFNO model incorporates a powerful vision transformer (ViT) backbone (Dosovitskiy et al., 2021), a neural network architecture originally designed for computer vision tasks. The ViT divides an input image into fixed-size non-overlapping patches (tokens) that are fed into the transformer. By combining ViT with a Fourier transform-based token-mixing scheme, AFNO enables training high-fidelity data-driven models (Pathak et al., 2022).

FourCastNet is trained on the ERA5 dataset, which was already explained in Section 3.1.1, and in terms of practical use, it offers a notable advantage in computational efficiency. Although training FourCastNet requires an energy investment comparable to running a 10-day forecast with 50 ensemble members in the Integrated Forecast System (IFS), generating a forecast with FourCastNet is more efficient, requiring about 12,000 times less energy to generate a forecast than the IFS model (Pathak et al., 2022). This makes it feasible to generate large ensemble forecasts at a relatively low computational cost.

The FourCastNet Version 2 (V2) is used in this work, which employs Spherical Fourier Neural Operators (SFNO) for improved representation of nonlinear atmospheric dynamics. Instead

of using flat Euclidean domains, which can handle global poles inaccurately, FourCastNet V2 employs a Spherical Harmonic Transform (SHT), designed for spherical geometries, to enable more accurate and realistic simulations of global atmospheric dynamics (Zhou, 2024). Additionally, FourCastNet V2 is trained with denser vertical data, encompassing 13 pressure levels, while maintaining the same spatial and temporal resolution as FourCastNet. Figure 3.1 depicts the overall structure of the SFNO network $F_\vartheta : \mathbf{u}_n \to \mathbf{u}_{n+1}$, which maps a discrete state vector $\mathbf{u}_n$ at a time t_n to the state $\mathbf{u}_{n+1}$ at time t_{n+1}. The SFNO architecture consists of three main parts: An encoder network, multiple SFNO blocks, and a decoder network. The encoder network uses Multi-Layer Perceptron (MLP) layers - fully connected neural networks - to *inflate* the channel dimension, projecting the input data into a higher-dimensional feature space. The decoder *deflates* the data back to the desired output dimensions. Between the encoder and decoder, the core SFNO blocks model the dynamics of the data. Additionally, a learned position embedding is included after the decoder layer to model spatial dependencies across the spherical domain.

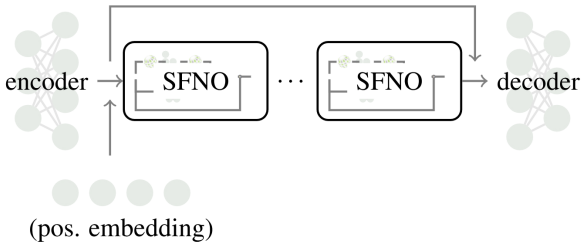


Figure 3.1: From Bonev et al. (2023) - Diagram of the overall SFNO architecture. Encoder and decoder MLPs inflate and deflate the channel dimension, and at the core of the architecture lie N SFNO blocks. A learned position embedding is added in cases where position-dependent information should be learned by the network.

3.2 The TEEMLEAP testbed

This section provides an overview of the TEEMLEAP testbed, used in this project. Developed by an interdisciplinary team from KIT and DWD, TEEMLEAP serves as a global framework for atmospheric prediction research. A general review of the testbed components is followed by a detailed explanation of the PSOs generation process through observation perturbation, concluding with an overview of the assimilation process within the Basic Cycling (BACY) Environment.

3.2.1 The TEEMLEAP testbed's prediction chain

Implemented on KIT's supercomputer HoreKa, the TEEMLEAP testbed emulates the entire operational weather forecasting chain in a simplified yet realistic form, enabling a systematic comparison of different configurations. At the same time, TEEMLEAP is sufficiently complex to allow extensions to relevant scales and real-world problems (Wilhelm et al., 2024). It enables the systematic investigation of key topics in weather prediction, including optimizing observational systems, quantifying uncertainty, and developing hybrid systems that integrate AI components with the existing physics-based models (Wilhelm et al., 2024).

An overview of the TEEMLEAP prediction chain is shown in Figure 3.2. First, ERA5 reanalysis data (see left panel of Figure 3.2) is used to draw PSOs, which are then perturbed to generate perturbed PSOs (symbolized by the red balloon in the figure, because they mimic radio-soundings). The perturbed PSOs are then input into DWD's BACY environment for assimilation-prediction-cycling experiments. These experiments use the Data Assimilation Coding Environment (DACE) and the ICON modeling framework (see the middle panel of Figure 3.2). More concretely, the TEEMLEAP testbed experiments are designed to allow: PSO-only, assimilation only, forecast-only or verification-only experiments; full cycling experiments comprising all components from the generation of PSO to verification; or experiments with combinations of several components. Additionally, it provides interfaces for incorporating new data-driven components, and integrating statistical and AI-based post-processing methods is also planned (Wilhelm et al., 2024).

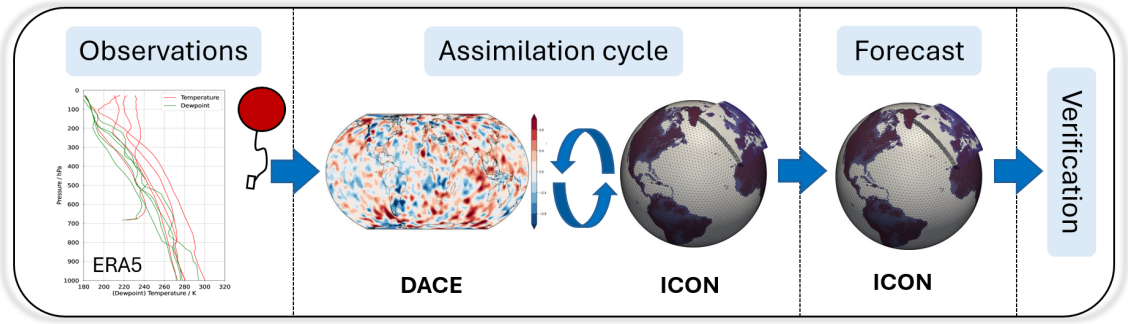


Figure 3.2: Overview of the prediction chain for the TEEMLEAP testbed. The left panel shows a vertical profile from the ERA5 reanalysis dataset. PSOs are derived from it and perturbed to generate perturbed PSO profiles, mimicking radio-soundings (symbolized by the red balloon). The middle panel represents the assimilation-prediction-cycling experiments performed using DWD's BaCY environment, which incorporates DACE and ICON. The panels on the right show the forward integration with ICON, using the output from the assimilation step, and the verification step.

3.2.2 Observation perturbation method

This section explains the generation of perturbed PSOs from ERA5 data for assimilation, following the method of Houtekamer et al. (1996), implemented in the MSC. This is the first step of the TEEMLEAP prediction chain represented in Figure 3.2.

First, ERA5 data is interpolated onto a Fibonacci lattice, a grid structure that distributes the interpolated stations almost uniformly and isotropically over the globe. The Fibonacci lattice is derived from the Fibonacci sequence, a series of numbers where each value is the sum of the two preceding numbers. Unlike the standard latitude-longitude grid, this structure avoids the issue of converging meridians near the poles and eliminates the need for Fourier filtering (Swinbank and James Purser, 2006). The resulting vertical profiles from this interpolation are referred to as PSO profiles. The number of PSO stations can vary from 0 to approximately O(10,000) (Wilhelm et al., 2024). Figure 3.3 provides two examples, showing configurations with 1,000 (panel a) and 2,000 (panel b) globally Fibonacci-distributed PSO stations.

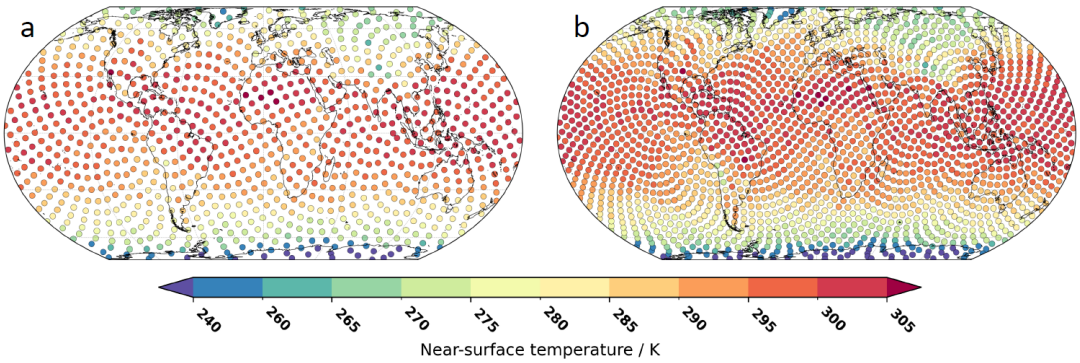


Figure 3.3: From Wilhelm et al. (2024), illustration example of the horizontal locations (dots) of the PSO stations on a Fibonacci lattice. The dots are colored according to the near-surface temperature on 30 September 2022 (00 UTC). The left panel (a) shows a configuration with 1,000 points, and the right panel (b) shows a configuration with 2,000 points. The Fibonacci lattice ensures a nearly uniform and isotropic distribution, which is a good approximation for locations equidistantly spaced.

For each vertical PSO profile, standard error profiles that simulate realistic uncertainties, are required prior to the assimilation, for the optimization procedure. These error profiles are necessary, as ERA5 values are not perfect. Without error profiles, observations would be treated as "perfect", receiving all the weight and ignoring the model first guess during the assimilation. In the testbed, this is achieved by diagnosing the typical ERA5-intrinsic error profiles using the method from Desroziers et al. (2005). These diagnostics are based on combinations of observation-minus-background, observation-minus-analysis, and background-minus-analysis differences, which provide a consistency check of an analysis scheme. The iterative procedure is conducted for several test cycling periods up to 1 month with O(1,000) globally distributed PSO stations (Wilhelm et al., 2024).

These diagnosed error profiles, called intrinsic error profiles, reflect the minimum uncertainty expected in observations. The profiles follow a Gaussian distribution centered at 0 with a standard deviation $\sigma_{ERA5}(p)$, which is defined globally. They embody three main sources of error: instrument error, which accounts for inaccuracies in the measurement process; representativeness error, which accounts for the subgrid-scale atmospheric phenomena not resolved by the model or the analysis; and observation operator errors, which originates from the transformation of model variables into observation space (Wilhelm et al., 2024; Kalnay, 2002). Concretely, only the fix/prescribed, diagonal random component of the error is considered. Adding them to the PSOs, the vertical profiles obtained are called "best-possible observations" or unperturbed PSOs, as they are considered the optimal estimate of the unknown true state of the atmosphere.

In scenarios where observational quality is assumed to be lower – i.e., with higher observation errors –, additional perturbations are added to the PSOs (Wilhelm et al., 2024). These perturbation errors increase the total observation error by introducing additional variability to the intrinsic errors. The standard deviation of the assumed global Gaussian-like distribution of these perturbation errors profiles is prescribed as follows (Houtekamer et al., 1996):

$$\sigma_{pert}(p) = \sqrt{f(p)^2 - 1} \cdot \sigma_{ERA5}(p) \tag{3.1}$$

in which $f(p)$ is a chosen pressure-dependent perturbation factor, that allows for the construction of arbitrary total error profiles. However, in this study, a constant value of f is used across all pressure levels.

Figure 3.4 shows an illustration of the perturbation profiles and their statistics. It is important to distinguish between the *perturbation profiles* (grey lines), which represent the individual vertical perturbation profiles for each PSO station; and *perturbation error standard deviation profile* (orange line), which is also a vertical profile, but reflects the standard deviation of the Gaussian-like distribution conformed by all the perturbation profiles at each pressure level. Additionally, the orange dotted line represents the theoretical standard deviation of the Gaussian distribution of perturbation profiles, assuming an infinite number of PSO stations. The figure also shows the standard deviation of the Gaussian-like distribution conformed by all the total error profiles (blue line), which account for both intrinsic and perturbation errors and is given by:

$$\sigma(p)^2 = \sigma_{ERA5}(p)^2 + \sigma_{pert}(p)^2, \tag{3.2}$$

and therefore:

$$\sigma(p) = f(p) \cdot \sigma_{ERA5}(p) \tag{3.3}$$

Figure 3.5 provides an illustrative sketch summarizing the connection between the defined errors. By setting the parameter f (constant across all pressure levels), and knowing the intrinsic error global statistics $\sigma_{ERA5}(p)$, then the perturbation error standard deviation profile $\sigma_{pert}(p)$ is derived. The globally distributed intrinsic errors at one single pressure level $p = p_j$, follow the gray-shaded Gaussian distribution with standard deviation $\sigma_{ERA5}(p_j)$; and the globally distributed perturbation errors at $p = p_j$, follow the blue-shaded distribution with standard deviation $\sigma_{pert}(p_j)$. The total errors, resulting from the combination of intrinsic and perturbation errors as defined by Equation 3.2.2, at a given pressure level $p = p_j$ follow the red-shaded Gaussian distribution on the right-hand side with a standard deviation of $\sigma(p_j)$.

The individual perturbation profiles are then constructed to satisfy the constraints previously discussed, while ensuring plausible and smooth profiles (Wilhelm et al., 2024). As mentioned, the default method in the TEEMLEAP testbed follows the approach from Houtekamer et al. (1996), using eigenvectors of a vertical covariance matrix. The sketch in Figure 3.5 also outlines the construction of the covariance matrix from the perturbed standard deviation profiles $\sigma_{pert}(p)$,

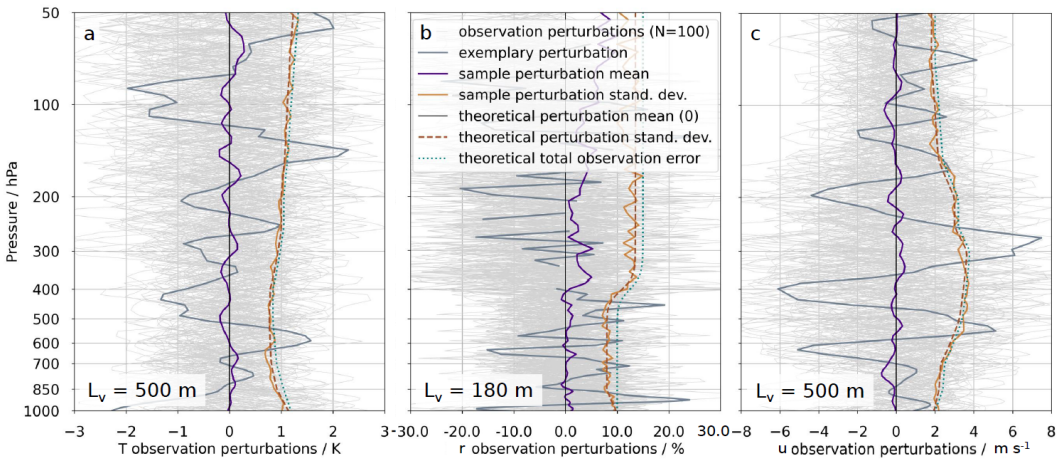


Figure 3.4: From Wilhelm et al. (2024), illustration of exemplary observation perturbation profiles for three meteorological variables: (a) temperature, (b) relative humidity and (c) zonal wind. The light grey lines depict N = 100 individual perturbation profiles generated, while the dark grey line represents a randomly chosen profile. The purple line shows the mean of the observation perturbations, and the orange line is the standard deviation of all the perturbation profiles. The ochre line shows the theoretical perturbation standard deviation, which reflects the standard deviation expected if there were an infinite number of PSO stations. The blue line indicates the theoretical total observation error profile, incorporating both perturbation errors and intrinsic errors from ERA5.

using the method from Houtekamer et al. (1996). The vertical covariance matrix is defined as the Hadamard product of a diagonal "perturbations matrix" and a correlation matrix, $\mathbb{P}$:

$$\mathbb{C} = [\sigma_{pert}(p_k)\sigma^T_{pert}(p_k)] \circ \mathbb{P} \tag{3.4}$$

The subscript k is introduced here to indicate that the error profiles are actually discretized across K levels rather than being continuous functions. This is because the intrinsic ERA5 error profiles are initially diagnosed on 14 significant pressure levels between 1000 hPa and 30 hPa, and then interpolated onto K = 1,030 levels between 1,030 hPa to 1 hPa spaced by 1 hPa, to obtain $\sigma_{ERA5}(p_k)$. The "perturbations matrix" is therefore a $K \times K$ diagonal matrix where each diagonal element corresponds to the square of the perturbation error standard deviation at the respective pressure level. On the other hand, the correlation matrix elements are defined by a Gaussian-shaped correlation function (Errico et al., 2013):

$$\rho(p_i, p_j) = \exp\left[-0.5\left(\frac{R_g T_o(\ln p_i - \ln p_j)}{gL_v}\right)^2\right] \tag{3.5}$$

where p_i and p_j are any two pressure levels. The diagonal elements, where $i = j$, are always equal to 1. In this equation, R_g is the gas constant, T_o is an approximate tropospheric mean temperature, and g is the acceleration of gravity. L_v is the vertical decorrelation length scale, which governs the correlation between different pressure levels. In the testbed, L_v is set to 500 m for wind and temperature, and 180 m for relative humidity (Errico et al., 2013), because relative humidity has smaller-scale variability in the troposphere due to cloud, precipitation and mixing processes

(Wilhelm et al., 2024). If $L_v \approx 0$, the correlation matrix would become diagonal, resulting in uncorrelated random sounding perturbations (Errico et al., 2013). However, with a non-diagonal correlation matrix (as in this case), vertically correlated perturbation profiles for each PSO station are obtained.

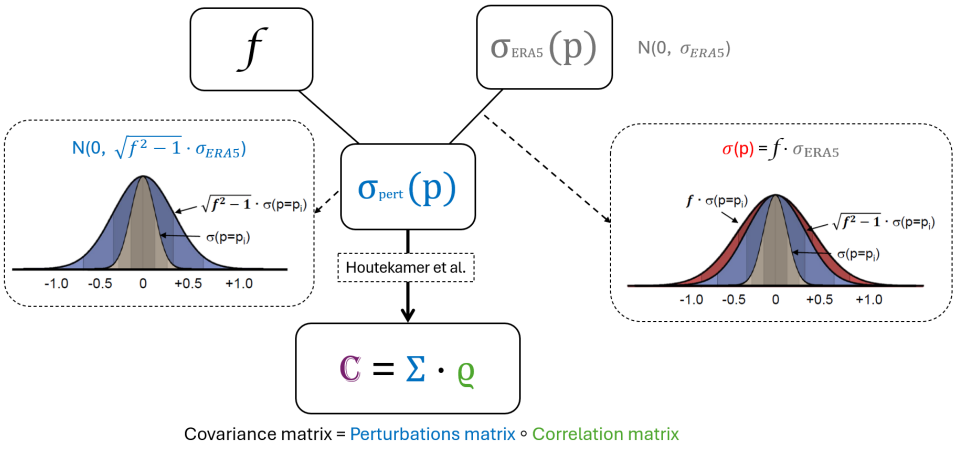


Figure 3.5: Schematic representation of the relationship between the different errors. The parameter f and the intrinsic error standard deviation $\sigma_{ERA5}(p)$ are used to derive the perturbation error standard deviation profile $\sigma_{pert}(p)$. The total error standard deviation is obtained as a combination of the intrinsic and perturbation errors. Gaussian distributions for each error type at a given pressure level $p = p_j$ are shown. At the bottom, the sketch illustrates how these perturbation statistics contribute to the construction of the covariance matrix, following the method of Houtekamer et al. (1996).

To generate individual perturbation profiles that account for vertical correlations while following the global statistics $\sigma_{pert}(p_k)$, an eigenvector analysis is performed on the vertical covariance matrix $\mathbb{C}$. This analysis results in K eigenvectors and eigenvalues. In our case, $K = 1030$, because, as mentioned, the perturbation error standard-deviation profiles $\sigma_{pert}(p_k)$ are generated for 1030 pressure levels.

For each variable and PSO station, random numbers w_k, drawn from a standard normal distribution $N(0,1)$, serve as weighting factors for constructing the perturbation profiles. These random numbers are independently generated using a predefined random seed, set with the NumPy's np.random.seed("seed") function. Once the seed is defined, each call to the random number generator advances the sequence, such that different sets of weighting factors w_k are obtained for each variable and station. Examples of the random seeds used will be provided in Section 3.3.2. The perturbation profiles are then obtained by multiplying the eigenvectors $\mathbf{v_k}$ with their corresponding eigenvalues λ_k, and modulating their contribution with the random weights w_k, following the method of Houtekamer et al. (1996). Mathematically, this is expressed as:

$$PP(p_k) = \sum_{k'=1}^{K} w_{k'} \lambda_{k'} \mathbf{v}_{k'}(p_k) \tag{3.6}$$

Here, k' indexes the eigenvectors, their associated eigenvalues, and the corresponding random weights, with a total of K modes, as determined by the size of the covariance matrix. Since each eigenvector $\mathbf{v}_{k'}$ has K components, the resulting individual perturbation profiles also $PP(p_k)$ span K discrete pressure levels p_k.

By adding both the intrinsic errors and the perturbation errors to the PSOs, the *perturbed* PSOs are obtained. However, a mapping process is required before adding the error: the generated error profiles on 1,030 pressure levels are remapped to the closest 98 pressure levels corresponding to the ERA5 model levels (and therefore, the PSOs levels), assuming a surface pressure of 1,013.25 hPa. This mapping process is necessary because PSO derived from ERA5 include observations on pressure levels that may vary spatially (Wilhelm et al., 2024).

Both the perturbed PSO profiles and their corresponding statistics, defined by the total error standard deviation profiles $\sigma(p_k)$ (from Equation (3.2.2)), are then used as inputs for the assimilation process. While the assumption of Gaussian-distributed errors seems valid for all variables, it is not appropriate for humidity, which is inherently not Gaussian-distributed (Wilhelm et al., 2024). Despite this, the testbed currently accepts this non-optimal use of humidity observations. To address the issue, any humidity values in the perturbed PSOs that are smaller than 0% or larger than 100% are eliminated before the assimilation.

3.2.3 BACY Cycling Environment

In the second step of the TEEMLEAP testbed, the ERA5-generated perturbed PSOs serve as inputs for the data assimilation process within the BACY Environment, illustrated in Figure 3.2 as the Assimilation cycle. The BACY Environment is a framework developed at DWD to manage the NWP chain, handling both data assimilation and forecasting. This environment is developed for experimentation and has been applied in numerous assimilation-related studies (Ruckstuhl and Janjić, 2020; Zeng et al., 2021; Reimann et al., 2023). The integration of perturbed PSOs is managed within BACY using the global DACE environment. For this study, DACE is configured for a 3-hour assimilation cycle, but it can be adjusted for 6-hour or 12-hour intervals. Each assimilation (ASS) cycle initiates a 3-hour forecast with the ICON modelling framework to generate the background state (first guess). The ASS cycles are started on 01 September 2022 (00 UTC) based on the operational ICON background from the DWD database interpolated to a 26 km horizontal resolution (R03B06) on 90 vertical levels. The cycling is done until 06 September 2022 (00 UTC), resulting in a total of 160 assimilation cycles.

The 3D-Var-based PSAS, introduced in Section 2.1, is applied to integrate observational data into the model (Wilhelm et al., 2024). The ICON background values are interpolated from the model grid onto the observation locations, corresponding to the PSO stations and the 98 vertical levels from ERA5. This mapping is performed internally in DACE, by applying the observation operator $\mathbf{H}$ to the ICON model's background fields. Then, the 3D-Var minimization problem is solved only for the observation locations, which is computationally more efficient than 3D-Var algorithms that operate in model space. The intermediate solution vector $\mathbf{w}$ is then mapped back from the

observation space to the model state space (grid), such that the analysis resides in the ICON model's state space at a resolution of 26 km on 90 vertical levels.

Based on the output from the ASS cycle period, BACY conducts a medium-range forecast (MAIN), for any lead time given by the user. The ICON model is primarily used for these MAIN forecasts; however, the TEEMLEAP testbed also provides interfaces for integrating new data-drive components (Wilhelm et al., 2024), as is the case in this study, as discussed in the next section. The verification (VERI) cycle follows, calculating the model equivalents of the observations (i.e., interpolating model values to match the time and location of real observations). This step evaluates the model's accuracy by directly comparing forecast outputs against observations (Wilhelm et al., 2024).

3.3 Extensions to the TEEMLEAP testbed

This section provides an overview of the experiments conducted in this project as an *extension* to the TEEMLEAP testbed, to explore different approaches for data-driven weather prediction within the TEEMLEAP testbed. It begins with a general review of the focus of the experiments, followed by an explanation of the process to generate an initial condition ensemble within the testbed, and the subsequent use of the deterministic data-driven model FourCastNet V2 to forward integrate the ensemble and obtain a probabilistic forecast.

3.3.1 Overview of the experiments performed

In this study, the TEEMLEAP testbed is utilized to explore ensemble prediction using perturbed initial conditions and the FourCastNet V2 model. This approach is illustrated in Figure 3.6, which is expanded from the Figure 1.1 presented in the introduction. For simplicity, the figure illustrates a 3-member ensemble, while the actual experiments performed employ 10-member ensembles. The experiments within the TEEMLEAP testbed focus on the following points:

- The observation perturbation method proposed by Houtekamer et al. (1996) and explained in Section 3.2.2, is applied to create five different 10-member ensembles of perturbed PSOs for data assimilation. Each ensemble is generated by applying one of the five different PSOs perturbation methods that will be detailed in Section 3.3.2. The left panel of Figure 3.6 represents one single ensemble of perturbed PSO profiles derived from the ERA5 reanalysis dataset (source data). For this ensemble, parallel assimilation cycles are performed using the ICON model's first guess (second panel), resulting in an optimized initial condition ensemble. The sequence: *Perturbed PSO ensemble* $\rightarrow$ *Parallel assimilation cycles* $\rightarrow$ *Initial condition ensemble* is repeated for each of the five perturbation methods, obtaining five different initial condition ensembles.

- The forward integration is performed using one of the initial condition ensembles, as will be detailed in Section 3.3.3. In this step, the deterministic, data-driven model FourCastNet

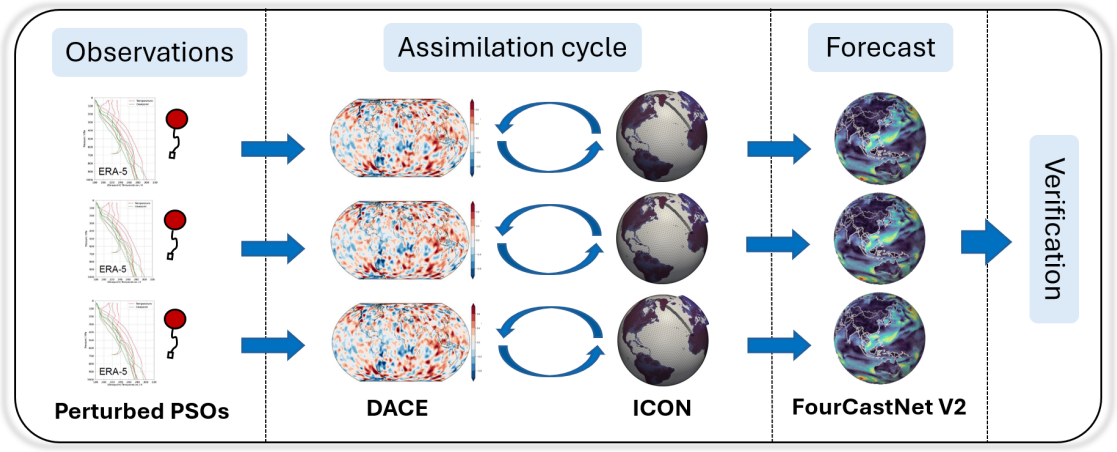


Figure 3.6: Overview of the experiments performed within the TEEMLEAP prediction chain expanded from Figure 1.1. For simplicity, the figure illustrates a 3-member ensemble, while the actual experiments performed utilize a 10-member ensemble. The left panel shows an ensemble of vertical profiles derived from the ERA5 reanalysis dataset (source data), representing perturbed PSOs designed to mimic radiosondes (symbolized by the red balloon). The second panel illustrates the parallel assimilation-prediction-cycling experiments performed for each perturbed observation set within the 10-member ensemble, using DWD's BaCY environment, which incorporates DACE and ICON. The third panel represents the forward integration of each ensemble member's analysis using the data-driven FourCastNet V2 model. The last panel represents the verification step of the prediction chain.

V2 is tested, performing parallel forward integrations separately from the data assimilation process with the NWP-based model ICON. Thus, a data-driven model is used instead of the default NWP-based model ICON model in the TEEMLEAP testbed. This step produces a data-driven probabilistic forecast, and is represented in the third panel of Figure 3.6.

3.3.2 Generation of initial condition ensembles within the testbed

Section 3.2.2 described the procedure for generating one set of globally distributed perturbed PSO profiles, from ERA5 data. However, to construct an initial condition ensemble for generating a probabilistic forecast, an ensemble of perturbed PSOs is required. To achieve this, the parameters for the construction of perturbed PSOs are differently modified.

In this study, five methods for differently perturbing the PSOs are tested. Each method focuses on certain parameters or their combinations to generate a 10-member ensemble of globally distributed perturbed PSOs (in total, five ensembles are generated). The details of each method are discussed below, with explanations that refer to the abstract 2D sketch in Figure 3.7 for better understanding. Table A in the Appendix provides a summary of the parameters used in each experiment.

Changing-seed: In this method the random seed is varied for each of the 10 members of the ensemble of perturbed PSOs. This ensures that the sets of random weighting factors - which determine the contribution of each eigenvalue and eigenvector in generating individual vertical profiles - differ across the 10 independently perturbed PSO files. However, within each member, the seed is the same across the assimilation steps, ensuring that the weighting

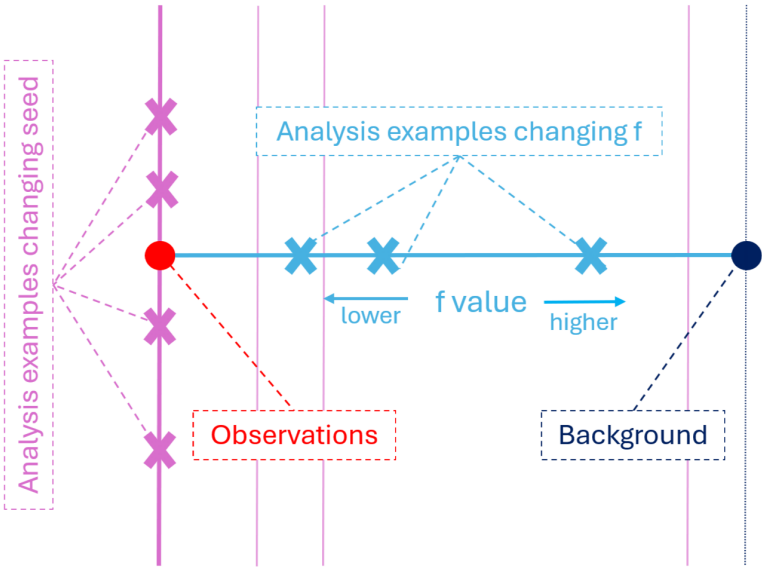


Figure 3.7: Abstract 2D sketch illustrating the perturbation methods for generating ensembles of perturbed PSO. The pink lines represent analyses resulting from the method of changing-seed perturbed observations, while keeping the f parameter constant (the bold pink line corresponds to $f = 1$, representing perfect observation quality). The light blue line corresponds to analyses obtained by varying the f parameter while keeping the seed constant. The combined method of changing both the seed and f parameter explores the entire space enclosed by the pink line and the dark blue dotted line, representing a broader range of analyses.

factors for each individual PSO profile do not vary throughout the assimilation process. Examples of seed values used include 777 for Experiment 1, and 247229 for Experiment 2, while the complete list of values can be found in Table A in the Appendix. Meanwhile, the f parameter is kept constant with a value of 2.0, such that the observational error during the assimilation is the same for all the ensemble members.

In the abstract sketch (Figure 3.7), this method corresponds to moving along the pink line, maintaining a constant "distance" to the background, as the observational error (determined by f value) remains unchanged. The thin light pink lines indicate that the position of the changing-seed resulting analysis depends on the chosen f value. The bold pink line represents the case where $f = 1$, meaning perfect observation quality, where observations are given full weight in the assimilation. In the case of this study, with $f = 2$, the pink line is placed closer to the center of the sketch.

This approach mirrors the method used by Houtekamer et al. (1996) and the CMC for generating an analysis ensemble, in which different random values are applied to the eigenvalues and eigenvectors to construct the perturbation profiles. As a result, the PSOs are perturbed in agreement with their error statistics, and it is the random variation between members that contributes to creating ensemble spread.

Changing-*f*: The f parameter, which governs the global statistics of total and perturbation error profiles, is varied for each of the 10 members of the ensemble of PSOs. In the assimilation experiments for this study, the f parameter ranges from 1.0 to 3.5 in 0.25 increments, with

each ensemble member assigned a different value. While the random seed remains constant - ensuring that the individual vertical profiles are weighted the same way across the ensemble - the variation in f changes the global observation error statistics, and with it, the quality of the observations. A smaller f value corresponds to higher observational trust, while a larger f value indicates greater observational error (lower observational trust), thereby reducing the PSOs influence on the analysis. This method therefore accounts for different levels of observational error within the ensemble, contributing to creating spread among the ensemble members.

Varying the f parameter while keeping the seed constant results in an analysis placed along the light blue line in the sketch shown in Figure 3.7. The exact abstract position of the analysis depends on the value of f: lower f values place the analysis closer to the observations (red point), while higher f values place it nearer to the background (dark blue point). For this study, the lowest f value is 1.0, and the highest is 3.5.

Changing-seed-*f*: In this perturbation method, a different seed value is set for each ensemble member, together with a varying f parameter that ranges from 1.0 to 3.5 in increments of 0.25. Therefore, the variation of both the observational error and the randomness in the construction of each individual perturbation profile, contribute to create spread.

In the abstract sketch shown in Figure 3.7, this method corresponds to moving across the entire space enclosed by the pink line and the dark blue dotted line. This represents a broader range of analyses within the ensemble.

Changing-*n*: In this method, each of the 10 members of the ensemble of perturbed PSOs is generated using a different number of PSO stations, referred to as "n" for simplicity. The number for each set ranges from 200 to 2000, increasing in increments of 200. Consequently, the ensemble members with a larger n have a finer spatial resolution. This enhanced resolution allows for a more accurate representation and a higher sensitivity to synoptic-scale phenomena like Rossby waves or convective systems. Therefore, it is possible to have ensemble spread between ensemble members by changing the number of perturbed PSO stations and capturing different sensitivities to these phenomena.

This method is not represented in the sketch in Figure 3.7. However, in that abstract space, it would be conceptually equivalent to moving along a pink line, as all ensemble members would have the same weight for observations during the assimilation process.

Changing-seed-*f*-*n*: Each of the 10 members of the ensemble of perturbed PSOs is created with a different seed for the individual profiles generation, a different value for the f parameter (ranging from 1.0 to 3.5 in 0.25 increments), and a different n (ranging from 200 to 2000 in 200 increments).

This approach corresponds to moving across the entire space enclosed by the pink line and the dark blue dotted line in the sketch. Additionally, the variation in n introduced an

extra degree of freedom among ensemble members, accounting for different resolutions and representations of atmospheric features across the ensemble.

3.3.3 Data-driven ensemble prediction using FourCastNet V2

For each of the five observation perturbation methods outlined in Section 3.3.2 (summarized in Table A in the Appendix), a 10-member perturbed-PSO-ensemble is generated and used in parallel ASS cycles within the TEEMLEAP testbed. The details of the assimilation experiments performed in this study are discussed in Section 3.2.3. The assimilation period for one selected method is extended to 20 days, producing analyses from 1 September to 20 September 2022 (00 UTC) at 3-hour intervals.

In this study, instead of conducting subsequent MAIN experiments with BaCy using ICON, the outcome from the assimilation process of the selected observation perturbation method is used to separately initiate forward integrations with FourCastNet V2. These forecasts are then evaluated using probabilistic verification techniques. To achieve this, the initial conditions generated during the assimilation experiments are transformed into a format compatible with FourCastNet predictions using the *ai-models* tool written by Wilhelm et al. (2024).

The substitution of FourCastNet V2 in place of the MAIN experiments with ICON is shown in Figure 3.6. A total of 10 parallel forecasts are produced, forming the probabilistic forecast. More specifically, 10 deterministic FourCastNet V2 forecasts are initialized daily at 00 UTC from 6 September 2022 to 20 September 2022, each with a leadtime of 360 hours and data output every 6 hours, resulting in 60 time steps per forecast.

3.4 Evaluation of Ensembles with RMSE and Spread

Ensemble prediction systems often exhibit underdispersion, as they tend to overlook certain sources of uncertainty (Fortin et al., 2014). In this work, the reliability of the ensemble prediction system is evaluated by comparing the root-mean-square error (RMSE) of the ensemble mean to the average ensemble spread. Following the mathematical derivation from Fortin et al. (2014), the average ensemble spread will be computed as the square root of average ensemble variance.

3.4.1 RMSE and Spread calculation

The RMSE measures the mean difference between the forecast ensemble mean and the truth, providing a sense of the overall deviation of the forecast from the truth (Yu et al., 2014). It is defined as the square root of the average of the squared differences between the ensemble mean at each time t and the corresponding observation:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\overline{X}_t - y_t\right)^2} \tag{3.7}$$

where T represents the number of time steps (defined differently for the analysis and the forecast, as will be explained later), $\overline{X}_t$ is the ensemble mean of the R members $\mathbf{X}_t = \{X_{t,r},\ r = 1,...,R\}$, and y_t is a set of T observations made a different times t.

Second, an unbiased estimator for the ensemble spread is given by Equation 3.8. It represents uncertainty in the forecast ensemble, capturing the differences between individual ensemble members at each time step (Fortin et al., 2014):

$$(\overline{s_t^2})^{1/2} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{R-1}\sum_{r=1}^{R}\left(\overline{X}_t - X_{t,r}\right)^2} \tag{3.8}$$

By combining equations 3.7 and 3.8, it follows that for a finite number of cases T, the following relationship should be approximately verified, with a correction factor depending on the ensemble size R:

$$RMSE \approx \sqrt{\left(\frac{R+1}{R}\right)}(\overline{s_t^2})^{1/2} \tag{3.9}$$

The correction factor serves to better represent uncertainty in small ensembles, assuming exchangeability between each ensemble member and the observation, but not between ensemble members (Fortin et al., 2014). For sufficiently large ensemble sizes, the correction factor that depends on R vanishes, such that the equation simplifies to:

$$RMSE \approx (\overline{s_t^2})^{1/2} \tag{3.10}$$

This expression assumes an infinite number of ensemble members and does not account for ensemble size limitations. As an example, Figure 3.8, taken from (Fortin et al., 2014), shows the comparison of both equations, for the case of the Global Ensemble Prediction System (GEPS) (Gagnon et al., 2013) from the Environment and Climate Change Canada (ECCC). The mathematical derivation from Fortin et al. (2014) shows the improved representation of the spread when calculating it as the square root of the average ensemble variance, instead of as the average standard deviation of ensemble forecasts, as done in many other studies (Charron et al., 2010; Gagnon et al., 2013).

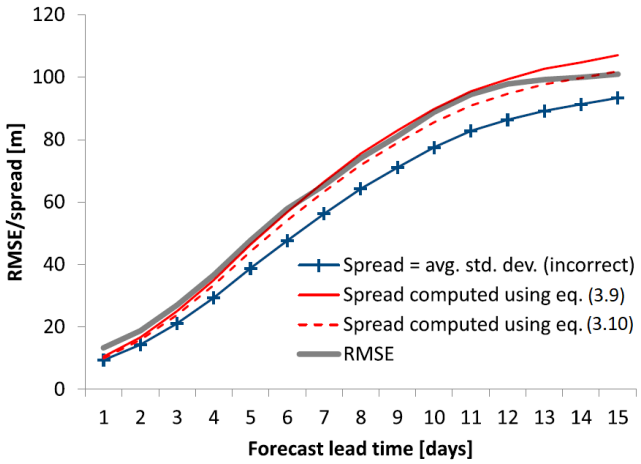


Figure 3.8: Plot taken from Fortin et al. (2014). RMSE and ensemble spread of the GEPS (Gagnon et al., 2013) from the ECCC, as evaluated in Fortin et al. (2014). The values are shown as a function of leadtime for 500-hPa height forecasts during the winter of 2011 (Fortin et al., 2014). This analysis exemplifies an RMSE/spread evaluation of a well-calibrated ensemble system, characterized by the alignment of RMSE and spread over time. It shows the comparison of Equation 3.9 (with correction factor) and Equation 3.10 (without correction factor for small ensembles).

3.4.2 Weighted global average of RMSE and Spread

In this study, both the weighted global average RMSE and the spread will be compared. To do so, the weighted global average is applied to both RMSE and spread values:

$$\overline{RMSE}_{\text{Global average}} = \frac{1}{N}\sum_{m=1}^{M}\sum_{n=1}^{N} L(m)\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\overline{X}_t[m,n] - y_t[m,n]\right)^2} \tag{3.11}$$

$$\overline{(\overline{s_t^2})^{1/2}}_{\text{Global average}} = \frac{1}{N}\sum_{m=1}^{M}\sum_{n=1}^{N} L(m)\sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{R-1}\sum_{r=1}^{R}\left(\overline{X}_t[m,n] - X_{t,r}[m,n]\right)^2}, \tag{3.12}$$

where the spatial location is denoted by the grid co-ordinates (m,n) and the weighting factor is:

$$L(m) = \frac{cos(m)}{\sum_{m=1}^{M} cos(m)}. \tag{3.13}$$

This leads to the following relationship:

$$\overline{RMSE}_{\text{Global average}} \approx \overline{(\overline{s_t^2})^{1/2}}_{\text{Global average}} \tag{3.14}$$

Or, with the correction factor for small ensembles:

$$\overline{RMSE}_{\text{Global average}} \approx \sqrt{\left(\frac{R+1}{R}\right)} \overline{(\overline{s_t^2})^{1/2}}_{\text{Global average}} \tag{3.15}$$

As a global average, a larger spread indicates greater general differences between the forecast charts (ensemble members), while a smaller spread suggests convergence among them. However, it is important to interpret spread with caution, as it does not reflect flow patterns but rather the uncertainty at each grid point (ECMWF, 2024b). For example, two similar looking forecast charts might have large differences (which translates in a large globally averaged spread) if they contain systems with strong gradients that are slightly out of phase. Conversely, two synoptically different charts might display a small overall spread if weak gradients dominate (ECMWF, 2024b).

Two different spread-skill relationships are evaluated: one for the analysis and one for the forecasts. For the analysis, the time mean of the RMSE over the entire cycling period T is computed, as shown in Equation 3.7, such that it reflects the overall error across the period, providing a stable measure of the performance of the assimilation system. On the other hand, the time mean is omitted from the spread calculation in Equation 3.8, resulting in a spread that varies with the cycling time step. As more observations are assimilated in each time step, the spread tends to increase slightly, because additional information is being integrated as the assimilation cycles progress.

For the forecast evaluation, the time mean is applied over all the initialization time steps for both the RMSE and the spread, so that the globally averaged values depend on the forecast lead time, providing an assessment of the forecast reliability.

3.4.3 Anomaly Correlation Coefficient (ACC) calculation

To evaluate the dispersion between two deterministic forecasts generated using different initial conditions, the latitude weighted ACC is used as a metric. The ACC reflects the pattern correlation between predicted and analyzed (truth) anomalies, highlighting errors in the position and strength of atmospheric flow regimes. Following Pathak et al. (2022), ACC for a certain variable, is defined as:

$$\text{ACC} = \frac{\sum_{m,n} L(m) \widetilde{X}_{pred,\,t}\,[m,n]}{\sqrt{\sum_{m,n} L(m) \left(\widetilde{X}_{pred,\,t}\,[m,n]\right)^2 \sum_{m,n} L(m) \left(\widetilde{X}_{true,\,t}\,[m,n]\right)^2}} \tag{3.16}$$

where $\widetilde{X}_{pred,\,t}$ is the anomaly of the forecast at a certain leadtime (t), calculated as the difference between the forecast prediction and the climatology; and $\widetilde{X}_{true,\,t}$ is the anomaly of the truth, calculated as the difference between the true state and the climatology.

4 Results

The main objectives of this study were outlined in the introduction. This section is organized to revisit each of those objectives, presenting the results obtained at every stage. The ultimate goal is to enable probabilistic forecasting using FourCastNet V2 within the TEEMLEAP testbed.

4.1 Evaluation of the ICON initial conditions from the TEEMLEAP testbed for running data-driven forecasts using FourCastNet V2

Since FourCastNet V2 is trained on ERA5 data produced by the ECMWF IFS model, it relies on initial conditions with a similar format to those of ERA5 to apply the SFNO network to predict their dynamics at later time steps (Pathak et al., 2022). However, in the TEEMLEAP testbed, the data assimilation process to produce initial conditions is performed by the ICON model from DWD. As a result, the output analysis differs in format from IFS ECMWF, making it incompatible with FourCastNet V2 predictions. Before generating forecasts using FourCastNet V2 in the TEEMLEAP testbed, a sanity check is required to confirm that ICON-generated initial conditions can be used and that FourCastNet V2 will produce reasonable forecasts based on these fields. To address this, the *ai-models* tool written by Wilhelm et al. (2024) is utilized to transform the ICON analysis obtained from the assimilation cycle of TEEMLEAP testbed experiments, into initial conditions readable by FourCastNet predictions. This analysis post-processing step supports the integration of data-driven models within the TEEMLEAP testbed.

The sanity check consists of generating initial conditions using the ICON model from the TEEMLEAP assimilation cycle. The assimilation is conducted using the PSAS system with a 13-km ICON background (R03B07) from the operational DWD archive and perturbed PSOs derived from ERA5. The PSO stations are located at 1,000 points uniformly distributed with 97 vertical levels, and slightly perturbed ($f = 2$) with vertical correlation. The generated initial conditions are then converted into a readable format by FourCastNet V2 using the *ai-models* tool. Then, a deterministic FourCastNet V2 forecast based on them is compared to a deterministic FourCastNet V2 forecast produced using the deterministic operational ECMWF IFS initial conditions.

The chosen period for this sanity check is selected based on the work from Zhou (2024) on the record-breaking 2021 Pacific Northwest Heat Wave, which occurred between 27 June and 1 July 2021. Figure 4.1 shows the global anomalies of 2 m temperature relative to the ERA5 30-year climatology (1990–2020). The figure includes subplots comparing forecasts initialized with IFS

and ICON initial conditions at three leadtimes: initial conditions (21 June 2021 – before the heat wave), 144 h leadtime (27 June 2021 – first day of the heat wave), and 312 h leadtime (4 July 2021 – after the heat wave event). The temperature anomalies at the initial conditions are nearly identical between the two datasets. Both forecasts successfully capture the heat wave occurrence on the 27 June 2021, as anomalously high values of the temperature in the Northwest American continent. Post-heat wave anomalies on the 4 July 2021 also exhibit consistent patterns between the two forecasts. While the general behavior across other parts of the world is very similar, some differences are noticeable, particularly at 312-hour leadtime in areas surrounding Antarctica and northern Russia. Nevertheless, these differences remain in a reasonable range of uncertainty for this long lead-time.

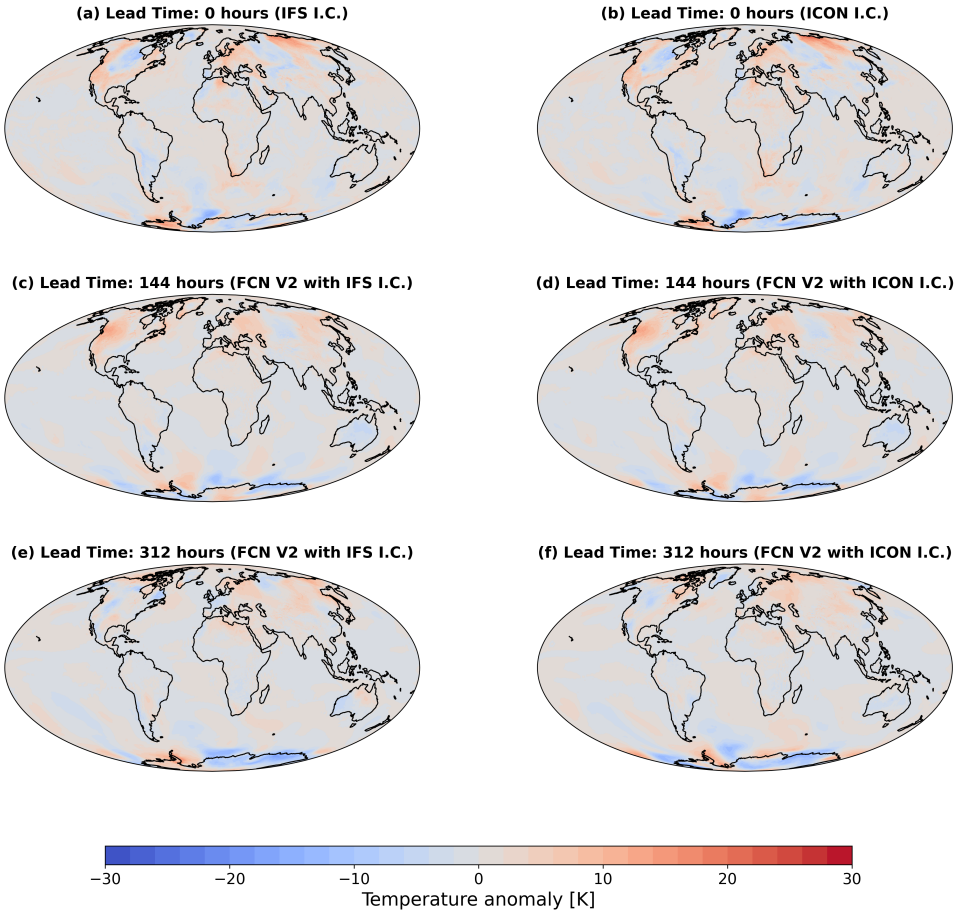


Figure 4.1: Global anomalies of 2 m temperature (K) for different leadtimes during the 2021 Pacific Northwest Heat Wave. Anomalies are computed with respect to the 30-year climatology from ERA5 (1990-2020). The left column ((a), (c), (e)) shows forecasts initialized with IFS initial conditions, and the right column ((b), (d), (f)) shows forecasts initialized with ICON initial conditions. Panels represent leadtimes of (a), (b) 0 h (21 June 2021, prior to the heat wave); (c), (d) 144 h (27 June 2021, first day of heat wave); and (e), (f) 312 h (4 July 2021, after the heat wave).

The applicability of ICON initial conditions generated from the TEEMLEAP testbed is further tested using the verification statistics of the ACC and the RMSE. Figure 4.2 presents the evolution with leadtime of these metrics (for the corresponding equations, refer to Section 3.4) of the forecasts for 500-hPa geopotential Z_{500} and 10 m zonal wind U_{10}. The left panels show the ACC, while the right panels depict the RMSE. Both metrics are computed as global averages to better reflect

the global behavior, as FourCastNet V2 is run globally. They are calculated using ERA5 data as the truth, and the climatology from ERA5 for the ACC. The results show that deterministic FourCastNet V2 forecasts, initialized with ICON initial conditions from the TEEMLEAP testbed, perform slightly better at longer leadtimes: they show higher ACC values and lower RMSE beyond 120 hours, compared to forecasts initialized with ECMWF IFS initial conditions.

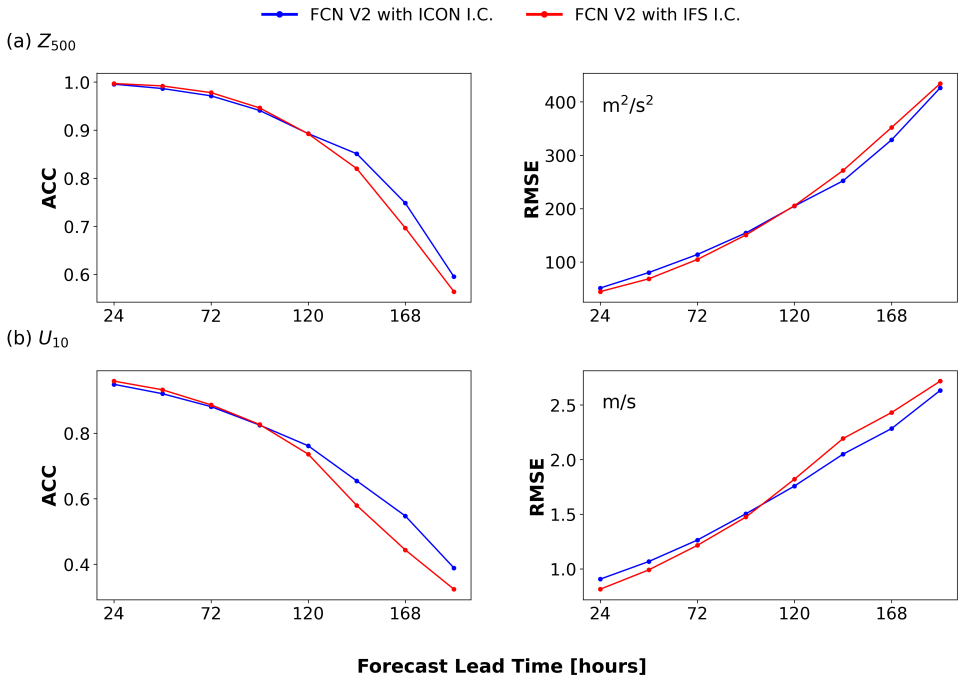


Figure 4.2: Evolution of ACC (left panels) and RMSE (right panels) as a function of forecast leadtime for (a, b) 500-hPa geopotential height Z_{500} and (c, d) 10-m wind speed (U10m). The blue line represents FourCastNet V2 forecasts initialized with ICON initial conditions, and the red line represents forecasts initialized with IFS initial conditions. Results are averaged globally, with ACC computed using anomalies from the 30-year ERA5 climatology (1990–2020) and using ERA5 analysis as the truth for both ACC and RMSE.

As RMSE here represents the global mean errors, regional variations might exist as it was seen in Figure 4.1, with ICON or IFS potentially performing somewhat better in specific areas. Nevertheless, the very similar evolution of ACC and RMSE in both forecasts confirms that ICON initial conditions are suitable for running deterministic FourCastNet V2 forecasts. This is the first successful use of ICON initial conditions generated in the TEEMLEAP testbed for running a pre-trained version of FourCastNet, serving as a sanity check that there are no technical issues and that forecasts are meteorologically plausible. This is the first step for implementing data-driven models for forward integration within the testbed, facilitated by the accessibility of the FourCastNet V2 code. Other data-driven models besides FourCastNet V2 can be easily integrated into the existing TEEMLEAP testbed structure, provided that they are open-source and ready-to-use. Notably, the DWD is now developing a data-driven weather prediction model based on data from its ICON model, called Artificial Intelligence ICON (AICON) (ECMWF, 2024a). This effort is part of Anemoi, a new community-based collaborative open-source initiative between a range of national meteorological services across Europe, including ECMWF (ECMWF, 2024a). It aims to enable meteorological organizations to train ML models with their own data. This highlights the interest in integrating ICON-based initial conditions into data-driven approaches.

4.2 Generation of ICON-based initial condition ensembles for data-driven ensemble prediction

4.2.1 Comparison of observation perturbation methods for initial condition ensemble generation

This section focuses on the production of an optimized initial condition ensemble by exploring different approaches based on the method from Houtekamer et al. (1996), introduced in Section 3.2.2. While they generated ensembles by randomly perturbing observations, this study compares three different ways of perturbing, as well as two combinations of them.

Assimilation experiments setup

By systematically varying the parameters in each of the observation perturbation methods, five 10-member perturbed PSO ensembles are created. The five different methods are: changing-seed, changing-f, changing-seed-f, changing-n, and changing-seed-f-n. For more details about the concrete parameter values used in each perturbation method, refer to Section 3.3.2. Once a perturbed PSOs ensemble is generated, its members serve as input for parallel data assimilation cycles using the ICON model within the TEEMLEAP testbed, resulting in an analysis ensemble. This procedure is repeated for each of the five perturbed PSO sets over a 5-day assimilation cycle period, spanning from 1 September 2022 (00 UTC) to 6 September 2022 (00 UTC), with assimilation performed at 3-hour intervals. As the first version of the TEEMLEAP testbed presented in Wilhelm et al. (2024) is employed, the same cycling period as they used is adopted.

The reliability of the resulting analysis ensemble is assessed using spread-skill statistics, by comparing the analysis spread (ensemble variability) with the RMSE. These results are crucial for constructing an initial condition ensemble, which is the foundation for generating a probabilistic forecast in the next step. High spread-skill correlations in the initial conditions of an ensemble system are desirable to start with, as they represent accuracy in the initial conditions and a good representation of the uncertainty. However, in this study, the forecasts will be performed using data-driven models. The connection between achieving a healthy spread-skill relationship in the initial conditions ensemble and maintaining it for the forecast ensemble may vary due to the different way these models work.

Observation perturbation methods: spread-error statistics in the initial condition ensemble

The following results focus on wind speed at 300 hPa WS_{300}. While the performance of the assimilation system was also investigated for the temperature at 850 hPa T_{850}, geopotential at 500 hPa Z_{500}, and specific humidity at 700 hPa Q_{700}, these results are not included here, as they exhibited similar patterns. Thus, WS_{300} is considered representative for evaluating the effectiveness of the analysis ensemble generation methods.

Figure 4.3 shows the evolution of the ensemble spread and RMSE, for each observation perturbation method over assimilation cycling time. The spread calculations for each ensemble are computed by sequentially increasing the number of ensemble members, highlighting how the spread changes when ensemble subsets with increasing number of members are considered. It was calculated using equation (3.10). Ideally, a correction factor for small ensembles should be included, but it has been omitted here to focus on the effect of increasing ensemble size. The correction factor will be included in later evaluations. On the other hand, to calculate the RMSE, the time mean over the entire assimilation period is taken. This provides a measure of the overall analysis accuracy across time, reflecting the time-mean deviation of the ensemble mean from the ERA5 truth. Since the system is updated every 3 hours with PSO profiles derived from ERA5, averaging the RMSE over time minimizes the influence of specific weather situations.

• Changing-seed, and changing-f

For the randomly changing-seed method (Figure 4.3a), the entire-ensemble spread (black line) stabilizes at approximately 2.3 $\mathrm{m\,s^{-1}}$ after about two days of cycling. Additionally, the spread collapses around 10 members, suggesting that this number of members is sufficient. In contrast, the changing-f method (Figure 4.3b) shows only slight variations in spread as more members are added. Although the f parameter was varied between 1 and 3.5, this range has little impact on spread. Additionally, changing-f shows the highest RMSE ($\sim$2.5 $\mathrm{m\,s^{-1}}$), and a very low spread ($\sim$1.25 $\mathrm{m\,s^{-1}}$), showing that this method does not represent the uncertainty of the ensemble.

• Changing-seed-f

In contrast, for changing-seed-f (Figure 4.3c), the spread increases more significantly as more members are added to the analysis ensemble evaluation, stabilizing near 13 members. This suggests that larger ensemble size would be required to effectively explore the performance of an ICON initial-condition ensemble using this method. Referring back to the sketch in Figure 3.7, the larger variation observed when varying both the seed and f together, compared to varying them separately, can be understood conceptually. It is analogous to exploring the entire space in the sketch, rather than moving along the lines, resulting in a broader and more diverse range of possible outcomes. This is likely also why the spread from changing-seed-f is slightly closer to RMSE compared to changing-seed (Figure 4.3a), suggesting a slightly better representation of uncertainty. However, the changing-f spread calculated considering the entire ensemble (black line) stabilizes after about three days of cycling, which is one day later than in the changing-seed plot.

• Changing-n, and changing-seed-f-n

When adding changes in the number of PSO stations (changing-n and changing-f, seed and n; Figures 4.3d and e, respectively), the spread increases substantially and deviates further from the RMSE, indicating a greater sensitivity to the ensemble size. It is important to clarify how the spread mean lines are constructed in those two subfigures. As described in Section 3.3.2, variations in n range between 200 and 2000. Spread Mean 1 corresponds to the ensembles of analyses created with 2000 and 1800 PSO stations. Spread Mean 2 includes the analyses performed with observation sets of 2000, 1800, and 1600 PSO stations, and this process continues in steps of 200 stations until

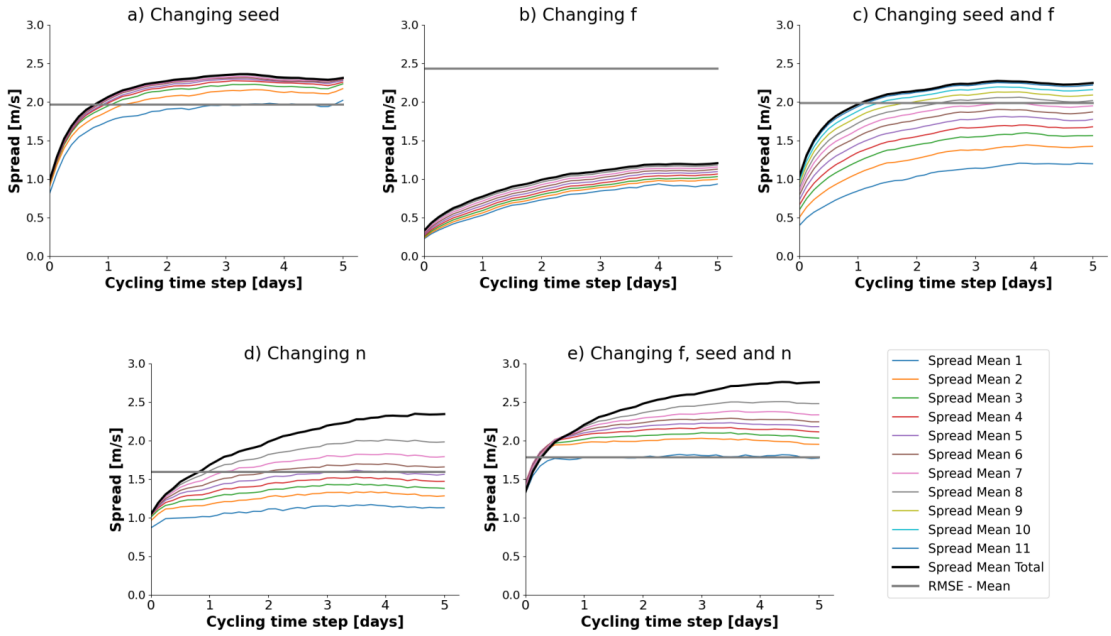


Figure 4.3: Spread and RMSE evolution of the analysis of WS_{300} for the five observations perturbation methods studied: (a) changing-seed, (b) changing-f, (c) changing-seed-f, (d) changing-n, and (e) changing-seed-f-n. The spread is calculated as the global mean of the root-square-difference between the ensemble mean (of a subset of members) and each member in that subset, for subsets including from 2 to all the members. Spread Mean 1 to Spread Mean 8 refer to calculations of the spread using sequentially increasing numbers of members. Spread Mean 9, 10 and 11 are specific to subplot (c), where three additional ensemble members were included. Spread Mean Total (black line) represents the spread calculated across the entire ensemble for each case. The RMSE (solid grey line) shows the time-mean RMSE for each ensemble.

the Spread Mean Total, that incorporates the entire range of analyses, from 2000 down to 200 PSO stations. Initially, this evaluation was conducted starting with the smallest n sets instead. In that case, Spread Mean 1 (which included $n = 200$ and $n = 400$ stations) was the highest and gradually decreased as members with larger n were added. However, the order was reversed (starting with the largest $n = 2000$) to better observe the evolution of the spread as more members are added. This highlights that the analysis sets obtained from assimilation with fewer PSO stations (e.g., 200, 400, and 600), exhibit greater variability and more pronounced differences compared to sets with higher numbers of PSO stations (e.g., 1600, 1800, and 2000 profiles). Furthermore, as seen in Figures 4.3d and e, the spread of the changing-n analysis ensembles is far from collapsing for an ensemble size of 10 members.

The methods that involve changing the number of PSO stations result in the lowest RMSE, with values below 2 m s^{-1}. This means that considering different sensitivities for the weather situation in the analysis ensemble members generates an ensemble mean that is closest to the truth. These methods also show a significant mismatch of the spread evolution with respect to the RMSE, as the spread continues to increase during the assimilation period. This opens an option for future studies: testing smaller ranges for the variation in the number of PSO stations. Although not included in the results, evaluation of ensembles with tighter ranges of n variations were conducted,

specifically: 7-member ensemble with n ranging from 400 to 1600, 5-member ensemble with n ranging from 600 to 1400, and 3-member ensemble with n ranging from 800 to 1200. These evaluations showed approximately the same RMSE, but a lower spread closer to the RMSE, indicating a possible improved performance. While the experiments performed did not include enough ensemble members to fully evaluate this behavior, it suggests a direction for future work.

Selected observation perturbation method to initiate the FourCastNet V2 forecast

The observation perturbation methods that perform best in terms of matching the spread and RMSE are the changing-seed method (Figure 4.3a) and the changing-seed-f method (Figure 4.3c), both achieving a time-mean RMSE of approximately 2 $\mathrm{m\,s^{-1}}$, and a spread of around 2.5 $\mathrm{m\,s^{-1}}$. Among these, the **changing-seed observation perturbation method** is selected to initiate forward integrations using FourCastNet V2, as the analysis ensemble spread stabilizes after two days of cycling, one day earlier than in the changing-seed-f method. This choice follows the principle of keeping-it-as-simple-as-possible, as the method, while not perfect, is already sufficient to generate an initial condition ensemble for the FourCastNet V2 forecast. While increasing the number of ensemble members could better represent initial condition uncertainty, focusing on a 10-member ensemble is sufficient to evaluate the performance of the observation perturbation approach. This ensemble size provides already a practical basis to try a probabilistic forecast within the TEEMLEAP testbed, while maintaining certain reliability in the results. However, in practical applications, one would eventually run large ensembles of data-driven models (50–1,000 members), due to their low computational cost.

In NWP systems, if the spread mismatches the RMSE during the assimilation cycle, this may propagate into the forecast, leading to errors in the representation of initial condition uncertainty (Houtekamer et al., 1996). However, in ensemble forecasts generated by data-driven models, their spread-skill performance have different sensitivity to initial conditions uncertainty compared to NWP systems. In this study, while achieving well-calibrated initial conditions would be desirable, it is not the primary focus. As a result, the initial condition ensemble spread is overestimated, which may lead to excessive forecast variability, reducing its reliability. If the changing-seed method is found inadequate for producing a well-balanced spread-skill relationship in the forecast, it can serve as a foundation for future investigations to explore and evaluate other perturbation methods or larger ensemble sizes. However, further exploration is beyond the scope of this study. The objective here is to test and demonstrate the feasibility of generating data-driven probabilistic forecasts through data assimilation with perturbed observations, serving as groundwork for future investigations, and validating the TEEMLEAP testbed's effectiveness for these experiments.

4.2.2 Evaluation of the initial condition ensemble generated with the changing-seed observation perturbation method

This section evaluates the performance of the changing-seed observation perturbation method in optimizing the initial conditions across four meteorological variables, for the subsequent forward integration using FourCastNet V2.

Insights from previous studies on the observation perturbation method

At this stage of the work, the focus is not yet on investigating the forecast performance; however, prior studies support the observation perturbation method for generating initial condition ensembles that lead to reliable forecasts. Hamill et al. (2000) compared different ensemble assimilation techniques, including SV, breeding methods, and the randomly perturbed observations method from Houtekamer et al. (1996). Their findings showed that, when assuming a perfect model, the observation perturbation method produced more accurate ensemble-mean analyses and subsequent better calibrated probabilistic forecasts, with higher spread-skill correlations compared to the other methods (Hamill et al., 2000).

In addition, Buizza et al. (2008) explored this method, including also stochastic physics to simulate the effect of model uncertainties (not included in this project) as described by (Houtekamer et al., 1996). They showed that using an EDA system based on the perturbed observations approach resulted in perturbations that were less localized geographically and provided better coverage of the tropics, compared to SV-based methods. However, it had too little spread compared to SV, due to a combination of too small initial amplitudes and too slow growth. They proposed integrating both methods for generating the initial conditions, to benefit from the strengths of both of them (Buizza et al., 2008). This work served the foundation for the ECMWF's current Ensemble Data Assimilation (EDA) system, operational since 2010 (Lang et al., 2019).

As in Hamill et al. (2000), in this work we do not explicitly account for model error. However, Houtekamer et al. (1996) noted that neglecting the model error could lead, over a number of assimilation cycles, to perturbations that are too small. To mitigate this, other studies explored approaches to address model error indirectly. For instance, in Buizza et al. (2003) they applied a multiplicative inflation factor of 1.8 to the resulting perturbations. This adjustment increased the ensemble spread in the initial conditions and compensated for an insufficient representation of model error during the assimilation (Buizza et al., 2003). Therefore, simulating model error by inflating the ensemble spread could be an option for future works. However, the current analysis shows that the ensemble spread is already higher than the RMSE in the initial conditions. This indicates that the resulting perturbations, while not perfect, are already sufficient to proceed with the forecast.

Spread-error statistics in the changing-seed initial condition ensemble

In Figure 4.4, the evolution of the spread and RMSE is shown for the changing-seed method applied to the four meteorological variables tested. The ensemble spread is evaluated using two equations from Fortin et al. (2014): equation (3.9), which includes a correction factor, and equation (3.10),

which omits it. The correction factor is added to better represent uncertainty in small ensembles, and in the case of this 10-member analysis ensemble, it results in a slightly higher spread. The spread-RMSE alignment is particularly good for T_{850} (Figure 4.4a), maintaining a stable spread-skill ratio which is 1.02 at the last time step. For the other variables, the spread remains generally stable but exceeds the RMSE, with spread-skill ratios of approximately 1.33 for Z_{500} (b), 1.14 for WS_{300} (c), and 1.25 for Q_{700} (d). The variable of Z_{500} behaves differently, maintaining an almost constant spread of approximately 0.4 dam, throughout the assimilation period. This is because, in the TEEMLEAP testbed version used for this study, the assimilation of geopotential is limited to the lowermost ERA5 model level, and higher levels, such as 500 hPa, were excluded. The minimal variation observed in Z_{500} reflects the indirect influence of other assimilated atmospheric variables, such as temperature or wind, on geopotential height.

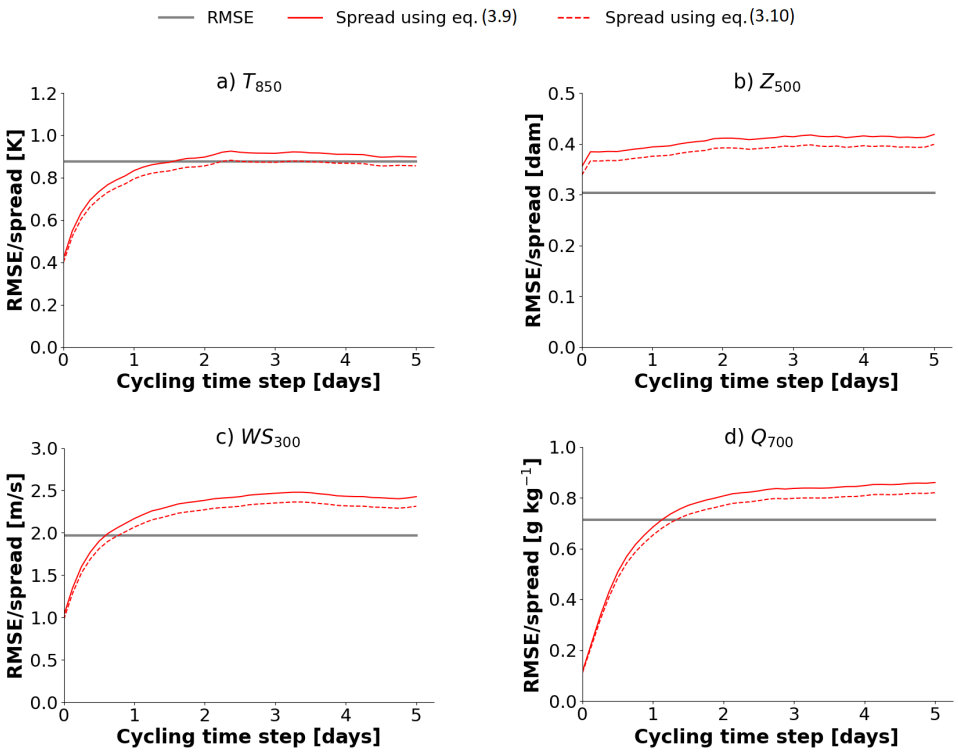


Figure 4.4: Spread (red lines) evolution, and time-mean RMSE (solid grey line) using the changing-seed observation perturbation method, for four different variables: (a) T_{850}, (b) Z_{500}, (c) WS_{300}, and (d) Q_{700}. The spread evolution during the assimilation period is shown for two equations: Equation (3.9) and Equation (3.10), as defined by Fortin et al. (2014) and detailed in Section 3.4.

Spatial differences in spread-error statistics in the Changing-seed initial condition ensemble

As indicated in Section 3.2.2, the error profiles follow global statistical properties. Provided that the number of PSO stations is high, then regional statistics will be similar to global statistics, meaning that limiting the assimilation to a specific region would not significantly disrupt these global properties. However, in the current version of the TEEMLEAP testbed only global experiments are supported (Wilhelm et al., 2024). Future and ongoing developments will enable limited domain experiments as well, but these were not explored in this study. Despite this, it is still valuable to

investigate how the current assimilation system performs across different regions. To this end, Figure 4.5 presents spread maps at the final analysis time step on 6 September 2022 (00 UTC), for the meteorological variables evaluated, and Figure 4.6 shows the RMSE and the spread evolution for each variable across three latitude bands.

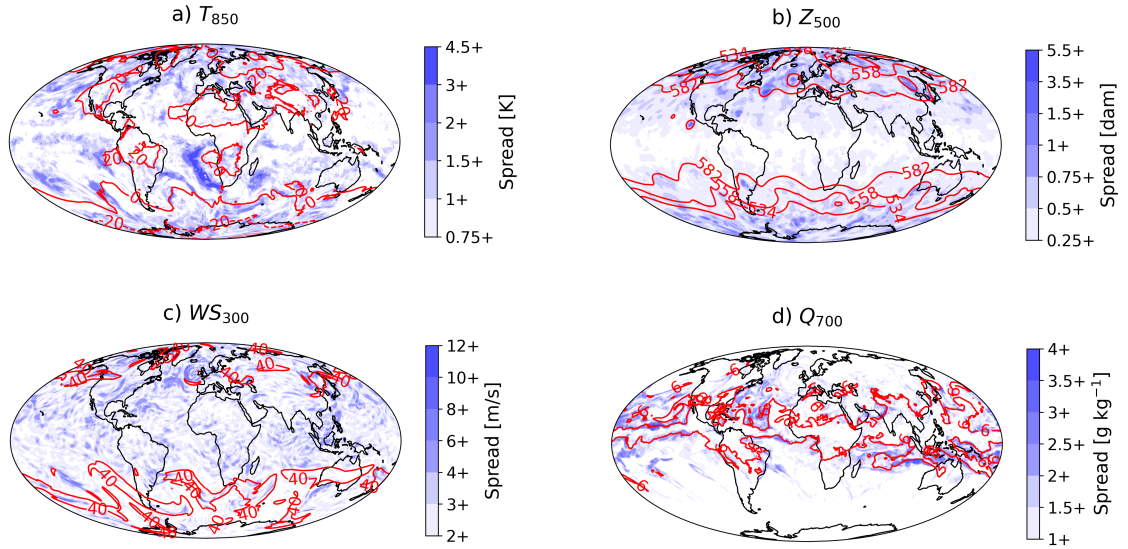


Figure 4.5: Maps of the spread at the final time step of the assimilation cycle, with ensemble mean average values depicted by red contours and spread indicated by purple shading. Plots are for the four evaluated atmospheric variables: (a) T_{850}, (b) Z_{500}, (c) WS_{300}, and (d) Q_{700}.

• 850-hPa Temperature

For T_{850} (Figure 4.5a), the spread is relatively high and constant, approximately 1.5 K, in both the Northern Hemisphere (NH) and Southern Hemisphere (SH) extratropics, while it is close to 0 K in the tropical regions. However, specific localized regions, particularly in the South Atlantic Ocean exhibit maximum values of approximately 4 K. Those localized areas appear only for T_{850} and not for other variables, suggesting a link to strong temperature inversions.

Next to the southwestern coast of Africa, the cold ocean surface driven by the Benguela current cools the air above it, creating a strong inversion with colder air trapped near the surface and warmer air on the top. This is enhanced by the presence of stratus clouds, which reduce surface heating and limit vertical mixing. As a result, the boundary layer in this region is estimated to be below the 850-hPa level. The sharp contrast between these stratified and adjacent well-mixed regions, leads to instability at these altitudes, amplifying the effect of small perturbations among the ensemble members. This leads to a higher spread in the analysis of T_{850}. Additionally, within the Intertropical Convergence Zone (ITCZ) core, the atmosphere is well mixed, but as it transitions to more stratified conditions in the surrounding regions, an increased spread is observed, particularly over the Pacific Ocean. These transition zones at the edges of the ITCZ, characterized by changes in stratification, contribute to localized variability and higher spread in the analysis of T_{850}. In contrast, over the Sahara region, the boundary layer reaches 700 hPa. This deep boundary layer ensures mixing of the atmosphere is well mixed with the surface, minimizing the impact of perturbations at 850 hPa and resulting in a low spread.

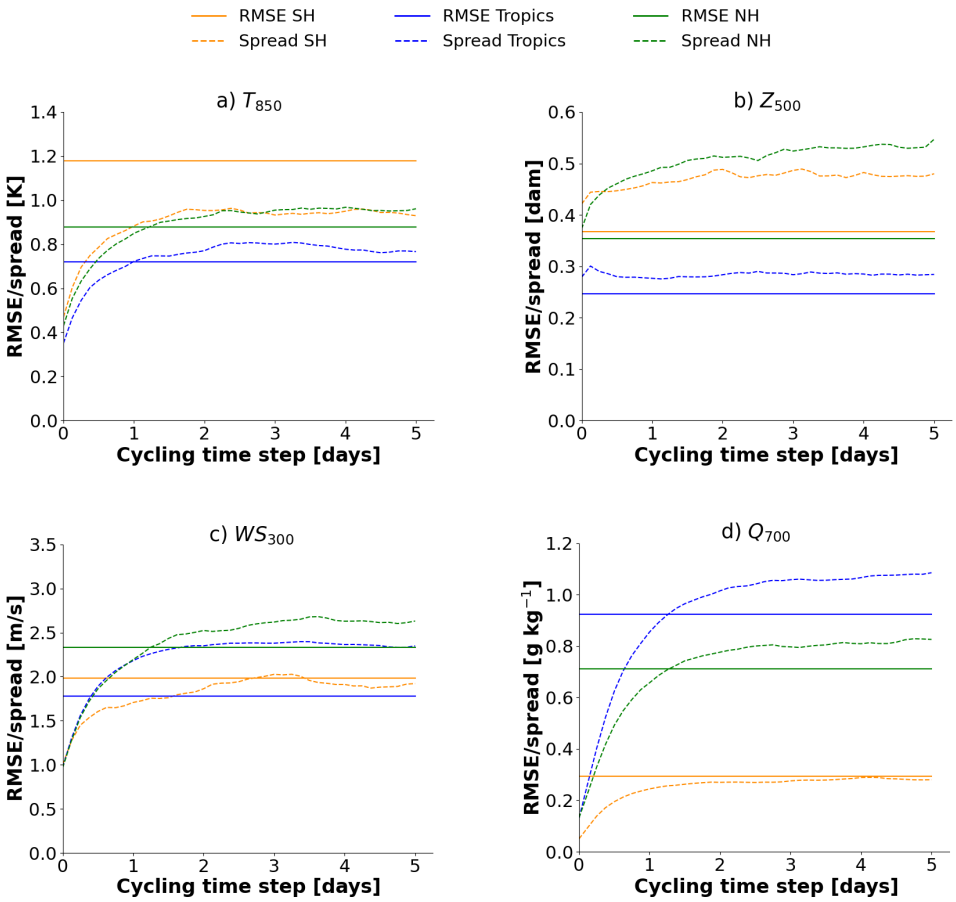


Figure 4.6: Time-mean RMSE and the evolution of the spread for the forecast leadtime, averaged across three latitudinal bands: NH extratropics (30ºN-90ºN), tropics (30ºS-30ºN), and SH extratropics (30ºS-90ºS). Plots are for the four evaluated atmospheric variables: (a) T_{850}, (b) Z_{500}, (c) WS_{300}, and (d) Q_{700}.

The T_{850} latitudinal averages (Figure 4.6a), show that the tropical regions exhibit the lowest spread and RMSE on average, and are approximately aligned. This is probably because of the lower temperature variability and the mixing from ITCZ. The SH extratropics, however, show the highest RMSE indicating that the ensemble mean deviates furthest from the truth, and also the spread underestimates the RMSE. The worse perfromance here could be attributed to the regions mentioned of transition to spring in the SH during September, which can increase temperature variability with more chaotic weather patterns and baroclinic instabilities.

• 500-hPa Geopotential

For Z_{500} (Figure 4.5b), the spread is higher in the NH and SH extratropical regions, with values of around 1-1.5 dam, compared to the tropics with around 0 dam. This indicates greater divergence among the ensemble members in the extratropics. The ensemble mean Z_{500} values, represented by the red contours, are lower and exhibit greater variation there, reflecting the higher atmospheric variability. In contrast, the tropics are encircled by a 582 dam contour line, indicating a more uniform ensemble mean, together with a lower spread. Figure 4.6b, shows that the spread closely

aligns with the RMSE in the spatially averaged values over the tropics. This could be related to the more stable and uniform weather structures in this region, where synoptic-scale variability is lower compared to the strong baroclinic activity that dominates the extratropical regions. However, as Z_{500} was not assimilated in this experiment, these characteristics result from residual effects of assimilating other atmospheric variables, and the geopotential in the lowermost ERA5 model levels.

• 300-hPa Wind Speed:

For WS_{300}, Figure 4.5c shows that, at the final analysis time step, the SH extratropics and polar regions exhibit higher ensemble mean wind speeds of around $40\,\mathrm{m\,s^{-1}}$, together with a spread smaller than $2\,\mathrm{m\,s^{-1}}$. This might be attributed to the end of the SH winter, with stronger and stable jet streams at high levels, reducing variability. In contrast, the spread in the tropics is higher, likely due to the influence of the ITCZ, where greater convection introduces variability in the horizontal wind speed at higher levels. The highest spatially averaged spread is observed in Figure 4.6c in NH extratropics. This can be explained by the presence of storm track regions, with strong westerly winds and high variability.

• 700-hPa Specific Humidity:

The spread map of the Q_{700} (Figure 4.5d) shows values smaller than $1\,\mathrm{g\,kg^{-1}}$ over the SH extratropics, where the ensemble mean values are also very low. This is expected, as the drier conditions in regions near the winter pole result in minimal Q_{700} variability, leading to low spread values. The spatially averaged spread (Figure 4.6d) in the NH extratropics is higher than in the SH, likely because NH is transitioning from summer, meaning that the atmosphere is warmer and more moist than SH. Lastly, the highest Q_{700} ensemble mean values found in the tropics (Figure 4.5d) are associated with the warm, moist air from convective activity. The blue regions at the margins of the ITCZ, indicate marginal stability in the central ITCZ, while the borders exhibit greater variability. This uncertainty is reflected in the higher RMSE and spread evolution observed in Figure 4.6d.

Summary

The analysis of spread and RMSE across the four evaluated atmospheric variables demonstrated that T_{850} has the closest spread-RMSE alignment, with higher variability in regions where there is a greater stratification. Additionally, it is evident that Z_{500} was not directly assimilated. Both WS_{300} and Q_{700} exhibit regions of high variability, associated with storm tracks in the NH mid-latitudes and convective activity in the tropics. In the SH extratropics, WS_{300} and Q_{700} exhibit lower variability, due to the presence of strong and stable winter jets, as well as drier conditions prevalent during winter.

As mentioned earlier, the assimilation in this study was performed on a global scale. Future developments of the TEEMLEAP testbed will enable limited-domain experiments, potentially improving the spread-error alignment in specific regions. However, this is beyond the scope of this study, which focuses on the global generation of perturbed PSOs, their assimilation with ICON (as discussed in this section), and the subsequent global probabilistic forecast using the data-driven model FourCastNet V2.

4.3 Evaluation of the FourCastNet V2 probabilistic data-driven forecast

The analysis ensemble was generated through assimilation with the ICON model within the TEEMLEAP testbed using the changing-seed perturbed PSOs. The optimized initial condition ensemble is now used to initiate forward integrations with FourCastNet V2.

4.3.1 Forecast setup

The FourCastNet V2 forecast performed utilizes initial conditions from 6-20 September 2022 at 00 UTC as starting points (in total, 15 starting points) for the simulations. This is done for each of the 10 analysis ensemble members generated with the changing-seed perturbation method. This means running 15 forecasts (for the 15 initialization times) for each of the 10 perturbed analysis of the ensemble $\rightarrow$ 150 forecast simulations in total. Each forecast is run for a leadtime of 360 hours (15 days), with outputs produced every 6 hours. It must be noted that the initial conditions from 1-6 September 2022 were excluded for the forecast, as they are considered a spin-up phase to ensure that the spread in the initial conditions took enough time to reach a stable value (refer to Figure 4.4). For the evaluations, RMSE and spread values are computed by time-averaging across all initialization times from 6 to 20 September 2022. The spread is calculated using equation (3.9), incorporating the correction factor to account for the small ensemble size of 10 members.

4.3.2 Spread-skill statistics of the FourCastNet V2 forecast

The ensemble spread is flow-dependent and varies for different meteorological variables (ECMWF, 2024b). Figure 4.7 displays the globally averaged results of the FourCastNet V2 probabilistic forecast for the four meteorological variables studied. The spread values at the initial conditions are 0.96 K for T_{850} (Figure 4.7a), 2.55 m s^{-1} for WS_{300} (Figure 4.7b), 0.43 dam for Z_{500} (Figure 4.7c), and 0.92 g kg^{-1} for Q_{700} (Figure 4.7d). As expected, these values closely match the spread of the analysis ensemble on 6 September 2022 (00 UTC) from Figure 4.4 in the previous section, which are 0.90 K, 2.42 m s^{-1}, 0.42 dam, and 0.86 g kg^{-1}.

A sharp decrease is observed at 1-day leadtime, with T_{850} and WS_{300} reducing to 58% (0.56 K) and 76% (1.95 m s^{-1}) of their initial spread. After this initial decrease, the spread increases steadily with forecast leadtime, reaching approximately 2.6 times (T_{850}) and 3.3 times (WS_{300}) the initial values by the end of the forecast period (15-days leadtime). A similar initial decrease is also observed in the Q_{700} forecast (Figure 4.7d), reaching 0.52 g kg^{-1} by day 3. In this case the impact is greater, as the spread recovers only slightly, surpassing the initial value at the end of the forecast (0.94 g kg^{-1} at 360 hours). A key factor for the system's difficulty in predicting Q_{700} could be inherently non-Gaussian nature of humidity errors, contrary to the assumption made within TEEMLEAP during the perturbation of observations to generate the initial conditions. In contrast, Z_{500} does not display a significant decrease in spread at shorter leadtimes, starting at 0.43 dam and decreasing

slightly to 0.41 dam after 6 hours. Beyond this point, the spread grows steadily, reaching a value 12.8 times larger than the initial conditions at the end of the forecast time.

The RMSE behaves more consistently across the four meteorological variables evaluated, increasing with leadtime. For T_{850}, WS_{300}, and Q_{700}, it stabilizes around day 10, with final values between 3 and 5 times larger than their respective initial RMSE values. Z_{500} behaves differently, with the RMSE continuing to grow, reaching by 15-days leadtime, a value 18 times larger than its initial value. This rapid increase of both RMSE and spread can be associated with the lack of assimilation for Z_{500}, which was discussed in the previous Section 4.2.2. As a result, the initial conditions were not directly constrained by observations and optimized, such that the forecast error diverges further and further from the truth, as the forecast progresses.

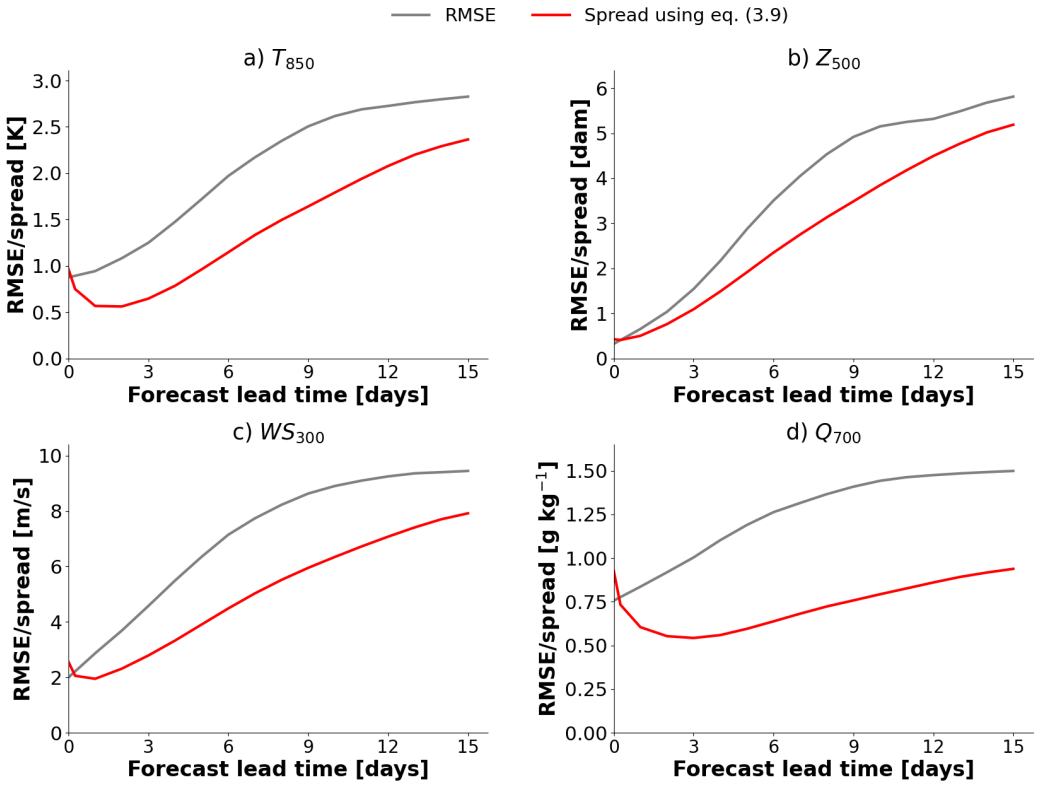


Figure 4.7: Globally averaged RMSE (gray line) and spread (red line) evolution for a leadtime of 15 days (360 hours) of the FourCastNet V2 probabilistic forecast for the variables: (a) T_{850}, (b) Z_{500}, (c) WS_{300}, and (d) Q_{700}. The spread is calculated using Equation (3.9) from Fortin et al. (2014).

The minimum spread-skill ratios occur around day 8 of the forecast, with the values shown in Table 4.1. These values are very close to the spread-skill ratio observed during the sharp decrease in spread on day 1 of the forecast. This indicates that the difference between spread and RMSE during the forecast is approximately proportional to the difference in the initial abrupt decrease. If there had been no abrupt decrease in spread for T_{850} and WS_{300} during the first 24 hours of the forecast, the spread and RMSE would align much more closely.

Table 4.1: Spread-skill ratios comparison: the minimum spread-skill ratio which occurs around day 8 of forecast, and the spread-skill ratio at the sharp decrease in day 1 of forecast.

Variable	Minimum ratio	Ratio at 24 h leadtime
T_{850}	0.65	0.60
Z_{500}	0.71	0.77
WS_{300}	0.67	0.68
Q_{700}	0.55	0.72

4.3.3 Spatial differences in spread-skill statistics of the FourCastNet V2 forecast

Given the behavior of the global data-driven forecast, it is valuable to examine error growth across different regions. To investigate this, Figure 4.8 shows the evolution of ensemble RMSE and spread for T_{850} as a representative variable. The spread maps in 4.8a show a decrease across all latitude bands during the first 2 days of forecast. It is particularly pronounced in the tropics, where the spread remains low for the rest of the forecast period. In contrast, the spread recovery in higher latitudes reflects a better performance of the model. The closest alignment between spread and RMSE is observed in the NH mid-latitudes. This improved performance could be related to the training of FourCastNet V2 with more accurate ERA5 data in this region.

At the initial forecast time, the highest spread values are observed over specific tropical ocean regions, particularly the equatorial Pacific and the South Atlantic. This matches the spread patterns seen at the analysis time step on 6 September 2022 (00 UTC) in Figure 4.5 (Section 4.2.2). Those maximum values are also reflected in the RMSE at 0-h leadtime, and they can be attributed to differences in stratifications caused by the Benguela current and the ITCZ, as previously discussed in Section 4.2.2. These features contribute to greater variability among ensemble members, leading to somewhat higher spatially averaged spread and RMSE in the SH tropics compared to the NH tropics (Figure 4.8d). Despite this, the tropics exhibit the lowest averaged T_{850} RMSE and spread throughout the entire forecast period, because of the lower temperature variability compared to higher latitudes.

In contrast, the highest spatially averaged RMSE and spread values occur in the polar regions, with higher values over the SH pole (Figure 4.8c) compared to the NH pole (Figure 4.8e). This can be attributed to the SH being the winter hemisphere during the forecast period, which increases temperature variability. Additionally, very high RMSE values are observed over Antarctica (Figure 4.8b), primarily because the 850-hPa pressure level is almost underground in this region due to the high topography. The steep sloping terrain and cold boundary layer outflows makes the Antarctic region difficult to be represented accurately in models (Connolley, 1996). This complexity may also contribute to the slight RMSE average decrease at the SH pole during the first forecast step.

A sudden decrease in RMSE is observed around days 10-11, in both the SH mid-latitudes (Figure 4.8c, blue lines) and the NH poles (Figure 4.8e, dark blue-green). This behavior in two geograph-

ically distinct regions suggests that the ensemble mean diverges from the truth in unexpected ways, raising questions about the system performance at extended leadtimes beyond 10 days. One possible explanation is that the forecast experiment is performed globally, which might lead to unexpected regional variations in performance. Another possibility is that the analysis is based on only 15 specific days of forecast initialization in September 2022, and this behavior might not occur when considering extended time ranges or forecasts initialized during other periods.

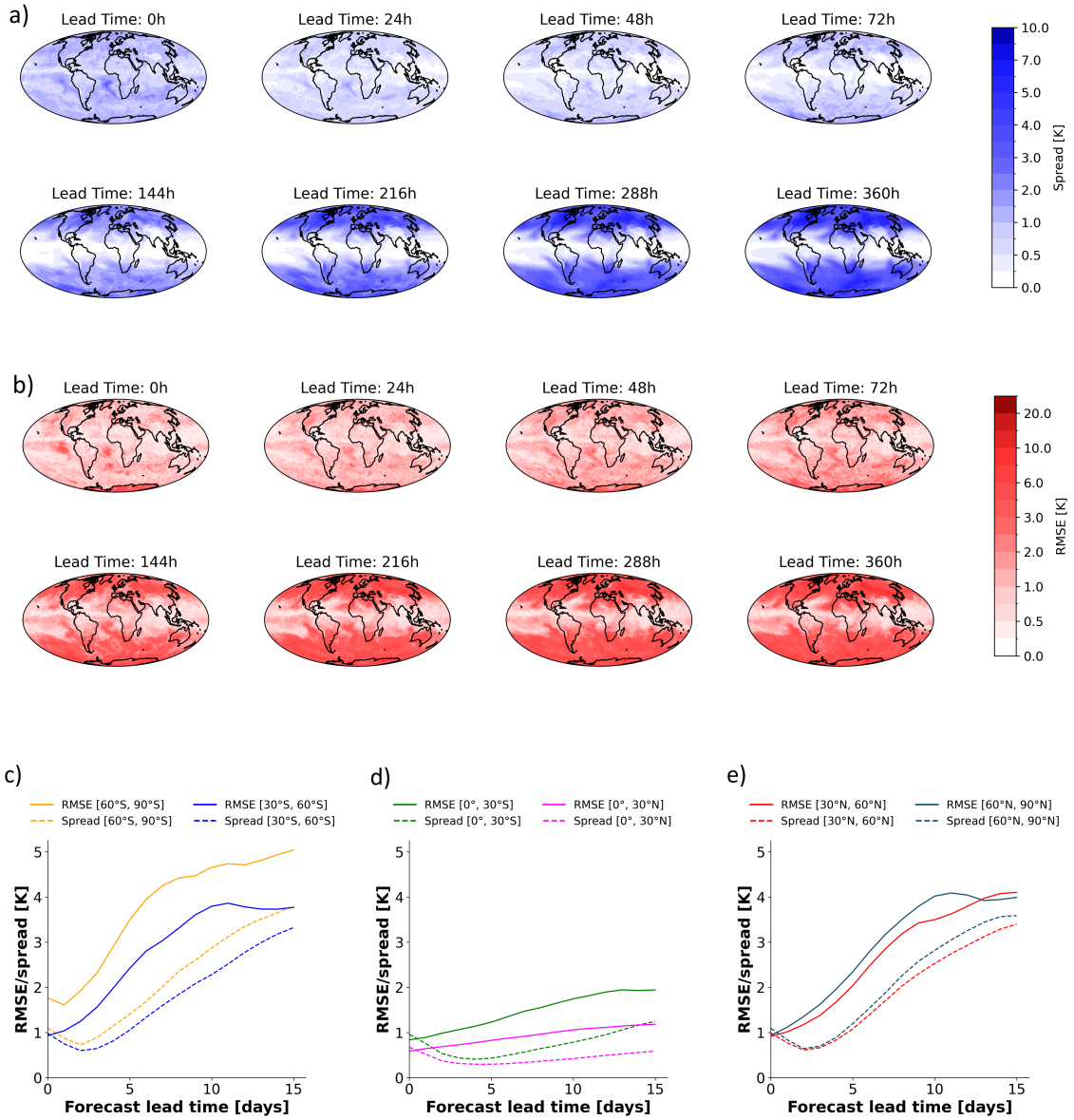


Figure 4.8: Evolution of RMSE and spread for T_{850} across different leadtimes, regions, and globally. Top rows (a) and (b): Global maps of spread (a) and RMSE (b) at different forecast leadtimes. Bottom figures are the spatially averaged RMSE and spread evolution over leadtime, for different latitude bands: (c) SH pole (60°S–90°S) and mid-latitudes (30°S–60°S); (d) SH tropics (0°–30°S) and NH tropics (0°–30°N); and (e) NH mid-latitudes (30°N–60°N) and pole (60°N–90°N). Solid lines represent RMSE, while dashed lines represent spread.

4.3.4 Discussion

Abrupt decrease in spread during the first 24 hours of forecast

In general, in NWP forecasts, the spread increases with forecast leadtime as uncertainties in the initial conditions propagate and grow (ECMWF, 2024b). However, in this data-driven forecast, a different error growth behavior is observed: the spread sharply decreases within the first 24 hours, making the forecast consistently underdispersive across all the variables tested. This decrease is evident both in the globally averaged evolution and across averaged latitude bands, indicating an abrupt shift in how complex structures are represented within the ensemble members during the first day of forecast. This is unlikely caused by a sudden transition in atmospheric conditions during September 2022, shifting from strong synoptic systems to larger-scale, fair weather systems with weaker gradients. Instead, the spread decrease is more likely related to the characteristics of the data-driven prediction system itself, and its difficulty to represent small-scale variability.

Bülte et al. (2024) found that for an ensemble forecast with ECMWF IFS operational initial conditions, using the data-driven model Pangu-Weather, the ensemble spread decreased during the first 24 h. Additionally, Bi et al. (2022) observed that the ensemble mean in Pangu-Weather forecasts sometimes introduced unexpected noise during the first days of forecasting. These findings suggest that ensembles using data-driven models may struggle to fully represent small-scale structures contained in the initial conditions, particularly during the first two days of forecast. Similarly, Pathak et al. (2022) generated an ensemble of 100 perturbed initial conditions with the Gaussian noise method and used it to produce probabilistic FourCastNet forecasts (AFNO-based model). They concluded that the model failed to account for several convective and radiative processes. To address these limitations, they suggested using an AFNO-based model trained with higher resolution data. In this study, the improved SFNO-based model FourCastNet V2 (see Section 3.1.3) is employed, but it is still insufficient to compensate for the limitations when maintaining and growing small-scale structures contained in the initial conditions.

Error growth in data-driven models

The system's limitations can be explained with error growth, as described by Baumgart et al. (2019) in their analysis using a stochastic convection scheme. They discussed that error growth in NWP follows a sequence of processes, with the first 12 hours dominated by differences in the convection scheme, primarily driven by latent heating and diabatic processes (Baumgart et al., 2019). However, as a data-driven model, FourCastNet V2 has fewer degrees of freedom compared to physical models, which are designed to represent small-scale features and parameterized processes contained in the initial conditions. As a result, it smooths out small-scale structures (Bi et al., 2022; Lam et al., 2023), failing to represent convective processes that dominate error growth during this stage. This likely explains the notable spread reduction in Q_{700}, as well as in T_{850} spread in tropical regions, as they strongly depend on small-scale processes like turbulent mixing and convection. The diffusion-based GenCast model from Google-DeepMind (Price et al., 2025), can address the fine-scale smoothing problem to some extent, but requiring a considerable increase in computational cost.

Between 12 h and 2 days, error growth is primarily driven by upper-tropospheric divergence, which projects errors from vertical-motion moist processes into the tropopause region (Baumgart et al., 2019). As errors upscale to larger scales, FourCastNet V2 begins to capture them, resulting in a recovery of the spread. This recovery occurs earlier for WS_{300}, which is nearer to the tropopause, while lower level variables, such as T_{850} and Q_{700}, are more influenced by surface processes, resulting in a somewhat slower recovery. After 2 days, error amplification up to the synoptic scale is dominated by differences in the nonlinear near tropopause dynamics (Baumgart et al., 2019), and therefore error growth becomes evident across all levels, as indicated by the increasing ensemble spread. However, since error growth is limited early in the forecast (Baumgart et al., 2019), variability among ensemble members is reduced before error growth at synoptic scales can take control, contributing to the low spread values observed.

The importance of model error

If all uncertainties associated with initial conditions and model errors are perfectly represented, the ensemble spread should equal the RMSE of the ensemble mean (Palmer, 2006). While this holds for NWP models, the forecast evaluated here is data-driven. As discussed earlier, FourCastNet V2's misrepresentation of small-scale features contained in the ICON-based initial conditions contributes to the spread-RMSE discrepancy, while the lack of model error representation may also play a role. Model error is believed to be an important component of the forecast error, particularly in the tropics (Reynolds et al., 1994). However, a limitation of all initial condition approaches for data-driven ensemble prediction is that they only account for initial condition uncertainty, and not for model uncertainty (Bülte et al., 2024). In contrast, physics-based NWP models incorporate techniques such as stochastic parameterizations to represent model error (Palmer, 2019b). Consequently, one limitation of the approach followed could lie in the exclusive focus on initial condition perturbations, without accounting for model error. As discussed in Section 4.2.2, neglecting model error can lead to perturbations that are too small (Houtekamer et al., 1996), and one possible way is to address this indirectly by applying an inflation factor (Buizza et al., 2003).

5 Conclusions

This thesis addresses the new wave of ML applications in weather prediction by investigating ensemble generation methods based on perturbing the initial conditions. While ensemble prediction with data-driven models research is still in its early stages, this study explored ensemble generation with perturbed observations. The experiments are conducted within the TEEMLEAP testbed, a prediction framework jointly developed by KIT and DWD, which mimics the operational weather prediction chain in a simplified but realistic manner, also incorporating interfaces for integrating AI-based components (Wilhelm et al., 2024).

The study extends the method proposed by Houtekamer et al. (1996) for ensemble generation by perturbing observations before data assimilation. Instead of relying only on vertically correlated random perturbations, as found in the literature, different alternative observation perturbation methods for ensemble generation were explored. While the new wave of data-driven weather research had primarily focused on deterministic forecasts, this study tested one of the ensemble generation methods discussed in Bülte et al. (2024) with the data-driven model FourCastNet V2. Finally, achieving data-driven ensemble generation, demonstrated TEEMLEAP testbed's value as a hands-on platform for students for exploring the weather prediction chain in educational contexts, and experimenting with ML applications.

This work addressed four key research objectives as outlined in the introduction:

1. Prove that the data-driven model FourCastNet V2 can be run with ICON-based initial conditions generated through the TEEMLEAP testbed instead of the ECMWF initial conditions it was trained with.
This was the first successful use of ICON-based initial conditions generated within the TEEMLEAP testbed for running a pre-trained version of FourCastNet. The forecasts generated were very similar to those produced using operational ECMWF IFS initial conditions, with ICON initial conditions even showing slightly better ACC and RMSE for longer lead times. This demonstrated the technical feasibility and ability to provide reasonable and meteorologically plausible forecasts.

2. Compare different methods for generating ensembles of initial conditions by modifying the parameters that control the nature of perturbed observations.
To create an ensemble of initial conditions, the method proposed by Houtekamer et al. (1996) was extended and refined. This involved systematically modifying observations before the data assimilation, exploring alternatives to the random perturbations found in the literature. Specifically, the following variations were tested: in the random seed used to generate the

vertical perturbation profiles, in the quality of observations between ensemble members, and the number of assimilated global observations. Additionally, combinations of these approaches were evaluated.

3. Select an adequate method for generating an initial condition ensemble with sufficient spread.
Among the tested observation perturbation methods to generate an initial condition ensemble, the changing-seed approach was selected to initiate forward integrations with FourCastNet V2. This choice was based on several criteria. First, it demonstrated one of the best alignments between RMSE and spread in the analysis ensemble, among the tested methods. Additionally, with 10 ensemble members, the spread stabilized without significant increases when more members were added. Moreover, during assimilation cycling, the spread collapsed more quickly than in other methods, eventually stabilizing at an approximately constant value. The method was also evaluated across four different meteorological variables, consistently showing good performance.

Following the "keep-it-as-simple-as-possible" principle, the changing-seed method was chosen, but it could be refined and extended in the future by more investigations. Possible extensions include increasing the number of ensemble members, introducing a longitude-dependent factor to the perturbation profiles, or testing the forward integration with other of the initial condition ensembles, such as with the one created by modifying the quality of observations before data assimilation.

Despite the possibilities, the procedure was not further refined in this study because the goal was not primarily to reach a perfect spread-skill relationship. In traditional NWP, a healthy spread-skill relationship in the initial conditions often means forecast quality. However, this may be different in data-driven models, and it is not yet clear how optimizing spread-skill in the initial conditions would impact the forecast. Therefore, the focus was on testing the feasibility and performance of the selected changing-seed perturbation method for generating an initial condition ensemble and its subsequent forward integration using a data-driven model.

4. Generate and evaluate an ensemble forecast from the perturbed ICON-based initial condition ensemble using FourCastNet V2 for the forward integration.
The combination of ICON initial conditions using the changing-seed observations perturbing method with the data-driven model for the 10-members ensemble forecast was found to be insufficient. The spread evolution showed an abrupt decrease during the first 24 hours of forecasting, particularly for lower-level variables. This indicates that during the first forward integration steps, the data-driven model smoothed out small-scale structures contained in the initial conditions, this way effectively making the ensemble members more similar to each other before error growth at synoptic scales could take control. After this, the spread increased again, evolving parallel to the RMSE but with a shift, resulting in a spread-skill ratio from approximately 0.60 to 0.70, depending on the variable.

The resulting ensemble forecast proved to be insufficient, indicating that the main goal of developing a sophisticated ensemble generation method for creating probabilistic data-driven forecasts within the TEEMLEAP testbed was not fully achieved. However, all the intermediate objectives leading to this goal were reached, and all the experiments were conducted within the TEEMLEAP testbed. This final step highlighted key differences in error growth characteristics between data-driven models and traditional NWP-based models, emphasizing the need for further research of data-driven models for ensemble prediction.

Since the combination of perturbed initial conditions based on ICON with FourCastNet V2 for forward integration was not entirely successful, future studies could explore using other data-driven models, such as PanguWeather. Further research could also focus on finetuning the perturbed observations. For instance, inflating the initial condition perturbations, as suggested by (Buizza et al., 2003), might help achieve a higher forecast spread by compensating for the lack of explicit model error representation. Additionally, alternative ensemble generation methods could be tested. For example, applying narrower ranges when varying the number of perturbed observation stations between ensemble members - such as 800 to 1200 stations instead of 200 to 2000 - could lead to improved spread-skill statistics. Another possibility is to optimize TEEMLEAP for regional experiments, and to perform limited domain assimilation and forecasting.

This study, within the TEEMLEAP framework, focuses on medium-range forecasting. However, there is an open question in MLWP research: Can data-driven models extend their predictive skill to subseasonal scales and climate prediction? This is a fundamentally different challenge, because, unlike weather forecasting, its evaluation metrics are not straightforward, and predictions at these longer timescales depend on different scenarios, such as greenhouse gas emissions or climate change mitigation strategies (Lee et al., 2021). A potential possibility to address subseasonal scales and climate prediction would be hybrid approaches. NeuralGCM (Kochkov et al., 2024), for example, combines a physics-based model with ML components, and accurately tracks climate metrics over multiple decades. A similar approach could therefore be explored by integrating the data-driven FourCastNet V2 with the physics-based model ICON. Hybrid systems would improve the representation of model error, for instance, by incorporating stochastic convection schemes to better represent convective processes. They could optimize the representation of small-scale structures in the initial conditions, advancing as well in subseasonal and climate prediction.

Finally, the error growth was not analyzed at different spatial scales, although it generally does vary depending on the scale, as shown in (Jung and Leutbecher, 2008). They analyzed error growth in ECMWF's EPS and found that during the first two days of the forecast, the ensemble was overdispersive due to excessive spread at synoptic scales. To address this, they reduced the amplitude of the initial perturbations. For longer forecasts, the ensemble became underdispersive at synoptic scales, which they attributed to insufficient spread at planetary scales, affecting the dynamics at smaller scales (Jung and Leutbecher, 2008). To solve this, they implemented a model that lead to larger perturbation growth. However, while (Jung and Leutbecher, 2008) analyzed error growth using an NWP-based ensemble, this study instead used the data-driven model FourCastNet V2. Since error growth behaves differently in data-driven models compared to traditional NWP

models, it would be interesting to investigate whether error growth in the data-driven ensemble system is more influenced by certain scales than others.

The ceiling for weather prediction with data-driven models is uncertain, raising questions about how far their skill can continue to improve. Currently, these models are trained on ERA5 reanalysis data, meaning their performance is linked to the quality and characteristics of ERA5 (Pathak et al., 2022; Bi et al., 2022; Price et al., 2025). This raises another question: How can observations be better incorporated into the training to improve the forecast? Frameworks like Anemoi are taking steps forward by enabling meteorological organizations to train machine learning models with their own data (ECMWF, 2024a). Such initiatives, together with many joint efforts, like the project TEEMLEAP, are improving weather prediction systems, and contributing to better extreme weather prediction for early warning systems (Field et al., 2012).

By exploring the relevance of AI weather research together with new methods for ensemble generation, this thesis represents a step forward in presenting the TEEMLEAP testbed as a valuable educational and research tool. Designing a first data-driven probabilistic forecast prototype with FourCastNet V2 provides an important first step into data-driven ensemble prediction with the TEEMLEAP testbed, bringing closer education and cutting-edge data-driven modeling. While there is still room for improvement and further development, the successful implementation of TEEMLEAP as a hands-on platform highlights its potential for educating the next generation of meteorologists and researchers.

6 Abbreviations

AFNO Adaptive Fourier Neural Operator

ACC Anomaly Correlation Coefficient

BACY BAsic-CYcling

CMC Canadian Meteorological Centre

DACE Data Assimilation Coding Environment

DL Deep-Learning

DNN Deep Neural Network

DWD Deutscher Wetterdienst, German Weather Service

ECMWF European Centre for Medium-Range Weather Forecasts

EDA Ensemble Data Assimilation

GCM General Circulation Model

GNN Graph Neural Network

HPC High-Performance Computing

ICON ICOsahedral Nonhydrostatic

IFS HRES High-RESolution configuration of the Integrated Forecasting System

ITCZ Intertropical Convergence Zone

ML Machine-Learning

MSC Meteorological Service of Canada

NH Northern Hemisphere

NN Neural Network

PSAS Physical-Space Assimilation System

PSO PSeudo-Observation

RMSE Root-Mean-Square-Error

SFNO Spherical Fourier Neural Operator

SH Southern Hemisphere

TEEMLEAP TEstbed for Exploring Machine LEarning in Atmospheric Prediction

A Appendix

Table A.1: Summary of the five methods used to generate ensembles of perturbed PSOs for probabilistic forecasting. Each method varies specific parameters (seed, f parameter, and/or the number of PSO n) while keeping others constant, as detailed in the table. The values for the varied parameters across the 10 ensemble members are listed for each method.

Method	Parameters varied: values = [Exp1,... Exp10]	Constant parameters
Changing-seed	Seed = [777, 247229, 4444, 4, 478902, 222, 490913, 128067, 19089, 632290]	f = 2.0 n = 1000
Changing-f	f = [1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, 3.0, 3.5]	Seed = 7718 n = 1000
Changing-seed-f	Seed = [7718, 230, 29831, 4590, 896, 81233, 9, 112415, 115, 34215, 35, 189] f = [1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, 3.0, 3.5, 3.75, 4.0]	n = 1000
Changing-n	n = [2000, 1800, 1600, 1400, 1200, 1000, 800, 600, 400, 200]	Seed = 3011 f = 2.0
Changing-seed-f-n	Seed = [915, 10, 1720, 980, 4, 360, 1234, 22, 1903204, 14] f = [1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, 3.0, 3.5] n = [2000, 1800, 1600, 1400, 1200, 1000, 800, 600, 400, 200]	-

Bibliography

Bauer, P., 2024: What if? numerical weather prediction at the crossroads. URL https://arxiv.org/abs/2407.03787, 2407.03787.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Baumgart, M., P. Ghinassi, V. Wirth, T. Selz, G. C. Craig, and M. Riemer, 2019: Quantitative view on the processes governing the upscale error growth up to the planetary scale using a stochastic convection scheme. *Monthly Weather Review*, **147 (5)**, 1713 – 1731, https://doi.org/10.1175/MWR-D-18-0292.1, URL https://journals.ametsoc.org/view/journals/mwre/147/5/mwr-d-18-0292.1.xml.

Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2022: Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. URL https://arxiv.org/abs/2211.02556, 2211.02556.

Bonev, B., T. Kurth, C. Hundt, J. Pathak, M. Baust, K. Kashinath, and A. Anandkumar, 2023: Spherical fourier neural operators: Learning stable dynamics on the sphere. URL https://arxiv.org/abs/2306.03838, 2306.03838.

Brunet, G., and Coauthors, 2023: Advancing weather and climate forecasting for our changing world. *Bulletin of the American Meteorological Society*, **104 (4)**, E909 – E927, https://doi.org/10.1175/BAMS-D-21-0262.1, URL https://journals.ametsoc.org/view/journals/bams/104/4/BAMS-D-21-0262.1.xml.

Buizza, R., 2008: The value of probabilistic prediction. *Atmospheric Science Letters*, **9 (2)**, 36–42, https://doi.org/https://doi.org/10.1002/asl.170, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/asl.170, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/asl.170.

Buizza, R., P. Houtekamer, Z. Tóth, G. Pellerin, M. Wei, and Y. Zhu, 2003: Assessment of the status of global ensemble prediction. *Mon Weather Rev*, **133**, 1076–1097.

Buizza, R., P. Houtekamer, Z. Tóth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ecmwf, msc, and ncep global ensemble prediction systems. *Monthly Weather Review - MON WEATHER REV*, **133**, https://doi.org/10.1175/MWR2905.1.

Buizza, R., M. Leutbecher, and L. Isaksen, 2008: Potential use of an ensemble of analyses in the ecmwf ensemble prediction system. *Quarterly Journal of the Royal Meteorological*

Society, **134 (637)**, 2051–2066, https://doi.org/https://doi.org/10.1002/qj.346, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.346, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.346.

Bülte, C., N. Horat, J. Quinting, and S. Lerch, 2024: Uncertainty quantification for data-driven weather models. URL https://arxiv.org/abs/2403.13458, 2403.13458.

Charron, M., G. Pellerin, L. Spacek, P. L. Houtekamer, N. Gagnon, H. L. Mitchell, and L. Michelin, 2010: Toward random sampling of model error in the canadian ensemble prediction system. *Monthly Weather Review*, **138 (5)**, 1877 – 1901, https://doi.org/10.1175/2009MWR3187.1, URL https://journals.ametsoc.org/view/journals/mwre/138/5/2009mwr3187.1.xml.

Chattopadhyay, A., P. Hassanzadeh, and D. Subramanian, 2020: Data-driven predictions of a multiscale lorenz 96 chaotic system using machine-learning methods: reservoir computing, artificial neural network, and long short-term memory network. *Nonlinear Processes in Geophysics*, **27 (3)**, 373–389.

Cohn, S. E., A. da Silva, J. Guo, M. Sienkiewicz, and D. Lamich, 1998: Assessing the effects of data selection with the dao physical-space statistical analysis system. *Monthly Weather Review*, **126 (11)**, 2913 – 2926, https://doi.org/10.1175/1520-0493(1998)126<2913:ATEODS>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/mwre/126/11/1520-0493_1998_126_2913_ateods_2.0.co_2.xml.

Connolley, W. M., 1996: The antarctic temperature inversion. *International Journal of Climatology*, **16 (12)**, 1333–1342, https://doi.org/https://doi.org/10.1002/(SICI)1097-0088(199612)16:12<1333::AID-JOC96>3.0.CO;2-6.

Desroziers, G., L. Berre, B. Chapnik, and P. Poli, 2005: Diagnosis of observation, background and analysis-error statistics in observation space. *Quarterly Journal of the Royal Meteorological Society*, **131 (613)**, 3385–3396, https://doi.org/https://doi.org/10.1256/qj.05.108, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/qj.05.108, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1256/qj.05.108.

Dosovitskiy, A., and Coauthors, 2021: An image is worth 16x16 words: Transformers for image recognition at scale. URL https://arxiv.org/abs/2010.11929, 2010.11929.

Düben, P., and Coauthors, 2021: 878. Machine learning at ecmwf: A roadmap for the next 10 years. ECMWF, URL https://www.ecmwf.int/node/19877, https://doi.org/10.21957/ge7ckgm.

ECMWF, 2023: The rise of machine learning in weather forecasting. URL https://www.ecmwf.int/en/about/media-centre/science-blog/2023/rise-machine-learning-weather-forecasting.

ECMWF, 2024a: Anemoi: A new framework for weather forecasting based on machine learning. https://www.ecmwf.int/en/about/media-centre/news/2024/anemoi-new-framework-weather-forecasting-based-machine-learning.

ECMWF, 2024b: Section 8.1.2 ens mean and spread. URL https://confluence.ecmwf.int/display/FUG/Section+8.1.2+ENS+Mean+and+Spread, accessed: 2024-12-31.

Errico, R. M., R. Yang, N. C. Privé, K.-S. Tai, R. Todling, M. E. Sienkiewicz, and J. Guo, 2013: Development and validation of observing-system simulation experiments at nasa's global modeling and assimilation office. *Quarterly Journal of the Royal Meteorological Society*, **139 (674)**, 1162–1178, https://doi.org/https://doi.org/10.1002/qj.2027, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2027, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.2027.

Field, C. B., and Coauthors, 2012: *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK, https://doi.org/10.1017/CBO9781139177245.

Fortin, V., M. Abaza, F. Anctil, and R. Turcotte, 2014: Why should ensemble spread match the rmse of the ensemble mean? *Journal of Hydrometeorology*, **15 (4)**, 1708 – 1713, https://doi.org/10.1175/JHM-D-14-0008.1, URL https://journals.ametsoc.org/view/journals/hydr/15/4/jhm-d-14-0008_1.xml.

Gagnon, N., and Coauthors, 2013: Improvements to the global ensemble prediction system (geps) from version 2.0.3 to version 3.0.0. Tech. note, Environment Canada.

Gneiting, T., and M. Katzfuss, 2014: Probabilistic forecasting. *Annual Review of Statistics and Its Application*, **1 (1)**, 125–151.

Goswami, P., K. C. Gouda, and O. Talagrand, 2005: Ensemble initial conditions through 4d-var assimilation. *Geophysical Research Letters*, **32 (21)**, https://doi.org/https://doi.org/10.1029/2005GL022542, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005GL022542, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2005GL022542.

Grover, A., A. Kapoor, and E. Horvitz, 2015: A deep hybrid model for weather forecasting. Association for Computing Machinery, New York, NY, USA, 379–386, KDD '15, https://doi.org/10.1145/2783258.2783275, URL https://doi.org/10.1145/2783258.2783275.

Guibas, J., M. Mardani, Z. Li, A. Tao, A. Anandkumar, and B. Catanzaro, 2022: Adaptive fourier neural operators: Efficient token mixers for transformers. URL https://arxiv.org/abs/2111.13587, 2111.13587.

Hamill, T. M., C. Snyder, and R. E. Morss, 2000: A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles. *Monthly Weather Review*, **128 (6)**, 1835 – 1851, https://doi.org/10.1175/1520-0493(2000)128<1835:ACOPFF>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/mwre/128/6/1520-0493_2000_128_1835_acopff_2.0.co_2.xml.

Hersbach, H., and Coauthors, 2023: Era5 monthly averaged data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), URL https://cds.climate.copernicus.eu, https://doi.org/10.24381/cds.f17050d7.

Ho, J., A. Jain, and P. Abbeel, 2020: Denoising diffusion probabilistic models. URL https://arxiv.org/abs/2006.11239, 2006.11239.

Houtekamer, P., and F. Zhang, 2016: Review of the ensemble kalman filter for atmospheric data assimilation. *Monthly Weather Review*, **144**, https://doi.org/10.1175/MWR-D-15-0440.1.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Monthly Weather Review*, **124 (6)**, 1225 – 1242, https://doi.org/10.1175/1520-0493(1996)124<1225:ASSATE>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/mwre/124/6/1520-0493_1996_124_1225_assate_2_0_co_2.xml.

Jung, T., and M. Leutbecher, 2008: Scale-dependent verification of ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, **134 (633)**, 973–984, https://doi.org/https://doi.org/10.1002/qj.255, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.255, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.255.

Kabri, S., T. Roith, D. Tenbrinck, and M. Burger, 2023: Resolution-invariant image classification based on fourier neural operators. URL https://arxiv.org/abs/2304.01227, 2304.01227.

Kalnay, E., 2002: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press.

Keresturi, E., Y. Wang, F. Meier, F. Weidle, C. Wittmann, and A. Atencia, 2019: Improving initial condition perturbations in a convection-permitting ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **145 (720)**, 993–1012, https://doi.org/https://doi.org/10.1002/qj.3473, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3473, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3473.

Kochkov, D., and Coauthors, 2024: Neural general circulation models for weather and climate. *Nature*, **632 (8027)**, 1060–1066.

Krizhevsky, A., I. Sutskever, and G. E. Hinton, 2012: Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., Curran Associates, Inc., Vol. 25, URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Lam, R., and Coauthors, 2023: Graphcast: Learning skillful medium-range global weather forecasting. URL https://arxiv.org/abs/2212.12794, 2212.12794.

Lang, S., E. Hólm, M. Bonavita, and Y. Tremolet, 2019: A 50-member ensemble of data assimilations. URL https://www.ecmwf.int/node/18883, 27-29 pp., https://doi.org/10.21957/nb251xc4sl.

Lee, J.-Y., and Coauthors, 2021: Future global climate: Scenario-based projections and near-term information. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, 553–672, https://doi.org/10.1017/9781009157896.006, URL https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_Chapter04.pdf.

Leutbecher, M., and T. Palmer, 2008: Ensemble forecasting. *Journal of Computational Physics*, **227 (7)**, 3515–3539, https://doi.org/https://doi.org/10.1016/j.jcp.2007.02.014, URL https://www.sciencedirect.com/science/article/pii/S0021999107000812, predicting weather, climate and extreme events.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, **20 (2)**, 130 – 141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/atsc/20/2/1520-0469_1963_020_0130_dnf_2_0_co_2.xml.

Lorenz, E. N., 1965: A study of the predictability of a 28-variable atmospheric model. *Tellus*, **17 (3)**, 321–333, https://doi.org/https://doi.org/10.1111/j.2153-3490.1965.tb01424.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2153-3490.1965.tb01424.x, https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2153-3490.1965.tb01424.x.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus A: Dynamic Meteorology and Oceanography*, https://doi.org/10.3402/tellusa.v21i3.10086.

Lynch, P., 2006: *The emergence of numerical weather prediction*. URL https://api.semanticscholar.org/CorpusID:134524393.

Magnusson, L., J. Nycander, and E. Källén, 2009: Flow-dependent versus flow-independent initial perturbations for ensemble prediction. *Tellus A*, **61 (2)**, 194–209, https://doi.org/https://doi.org/10.1111/j.1600-0870.2008.00385.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0870.2008.00385.x, https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1600-0870.2008.00385.x.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ecmwf ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, **122 (529)**, 73–119, https://doi.org/https://doi.org/10.1002/qj.49712252905, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49712252905, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.49712252905.

NVIDIA, 2023: Fourcastnet: Global weather forecasting with deep learning. URL https://build.nvidia.com/nvidia/fourcastnet.

Palmer, T., 2006: *Predictability of Weather and Climate*. Cambridge University Press, Cambridge, https://doi.org/10.1017/CBO9780511617652, URL https://doi.org/10.1017/CBO9780511617652.

Palmer, T., 2019a: The ecmwf ensemble prediction system: Looking back (more than) 25years and projecting forward 25years. *Quarterly Journal of the Royal Meteorological Society*, **145 (S1)**, 12–24, https://doi.org/https://doi.org/10.1002/qj.3383, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3383, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3383.

Palmer, T. N., 1993: Extended-range atmospheric prediction and the lorenz model. *Bulletin of the American Meteorological Society*, **74 (1)**, 49 – 66, https://doi.org/10.1175/1520-0477(1993)

074<0049:ERAPAT>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/bams/74/1/1520-0477_1993_074_0049_erapat_2_0_co_2.xml.

Palmer, T. N., 2019b: Stochastic weather and climate models. *Nature Reviews Physics*, **1 (7)**, 463–471, https://doi.org/10.1038/s42254-019-0062-2, URL https://doi.org/10.1038/s42254-019-0062-2.

Palmer, T. N., R. Buizza, R. Hagedorn, T. Jung, and M. Leutbecher, 2006: Ensemble prediction: A pedagogical perspective. *ECMWF Newsletter*, **106**, 10–17, URL https://www.ecmwf.int/sites/default/files/elibrary/2006/18024-ensemble-prediction-pedagogical-perspective.pdf.

Pasche, O. C., J. Wider, Z. Zhang, J. Zscheischler, and S. Engelke, 2025: Validating deep learning weather forecast models on recent high-impact extreme events. *Artificial Intelligence for the Earth Systems*, **4 (1)**, https://doi.org/10.1175/aies-d-24-0033.1, URL http://dx.doi.org/10.1175/AIES-D-24-0033.1.

Pathak, J., and Coauthors, 2022: Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. URL https://arxiv.org/abs/2202.11214, 2202.11214.

Poincaré, H., 1914: *Science and Method*. T. Nelson, London.

Price, I., and Coauthors, 2024: Gencast: Diffusion-based ensemble forecasting for medium-range weather. URL https://arxiv.org/abs/2312.15796, 2312.15796.

Price, I., and Coauthors, 2025: Probabilistic weather forecasting with machine learning. *Nature*, **637**, 84–90, https://doi.org/10.1038/s41586-024-08252-9, URL https://doi.org/10.1038/s41586-024-08252-9, published: 04 December 2024.

Rasp, S., M. S. Pritchard, and P. Gentine, 2018: Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, **115 (39)**, 9684–9689, https://doi.org/10.1073/pnas.1810286115, URL http://dx.doi.org/10.1073/pnas.1810286115.

Reimann, L., C. Simmer, and S. Trömel, 2023: Assimilation of 3d polarimetric microphysical retrievals in a convective-scale nwp system. *Atmospheric Chemistry and Physics*, **23 (22)**, 14 219–14 237, https://doi.org/10.5194/acp-23-14219-2023, URL https://acp.copernicus.org/articles/23/14219/2023/.

Reynolds, C. A., P. J. Webster, and E. Kalnay, 1994: Random error growth in nmc's global forecasts. *Monthly Weather Review*, **122 (6)**, 1281–1305, https://doi.org/10.1175/1520-0493(1994)122<1281:REGING>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/mwre/122/6/1520-0493_1994_122_1281_reging_2_0_co_2.xml.

Ruckstuhl, Y., and T. Janjić, 2020: Combined state-parameter estimation with the letkf for convective-scale weather forecasting. *Monthly Weather Review*, **148 (4)**, 1607 – 1628, https://doi.org/10.1175/MWR-D-19-0233.1, URL https://journals.ametsoc.org/view/journals/mwre/148/4/mwr-d-19-0233.1.xml.

Schultz, M. G., C. Betancourt, B. Gong, F. Kleinert, M. Langguth, L. H. Leufen, A. Mozaffari, and S. Stadtler, 2021: Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **379 (2194)**, 20200 097, https://doi.org/10.1098/rsta.2020.0097, URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2020.0097, https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2020.0097.

Soci, C., and Coauthors, 2024: The era5 global reanalysis from 1940 to 2022. *Quarterly Journal of the Royal Meteorological Society*, **150 (764)**, 4014–4048, https://doi.org/https://doi.org/10.1002/qj.4803, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.4803, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.4803.

Swinbank, R., and R. James Purser, 2006: Fibonacci grids: A novel approach to global modelling. *Quarterly Journal of the Royal Meteorological Society*, **132 (619)**, 1769–1793, https://doi.org/https://doi.org/10.1256/qj.05.227, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/qj.05.227, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1256/qj.05.227.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at ncep and the breeding method. *Monthly Weather Review*, **125 (12)**, 3297–3319.

Vannitsem, S., and Coauthors, 2021: Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bulletin of the American Meteorological Society*, **102 (3)**, E681–E699.

Warner, T. T., 2011: *Numerical Weather and Climate Prediction*. Cambridge University Press, Cambridge, UK, https://doi.org/10.1017/CBO9780511763243.

Weyn, J. A., D. R. Durran, R. Caruana, and N. Cresswell-Clay, 2021: Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *Journal of Advances in Modeling Earth Systems*, **13 (7)**, e2021MS002 502, https://doi.org/https://doi.org/10.1029/2021MS002502, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002502, e2021MS002502 2021MS002502, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021MS002502.

Wilhelm, J., and Coauthors, 2024: Teemleap - a new testbed for exploring machine learning in atmospheric prediction for research and education. https://doi.org/10.22541/essoar.173482059.96151727/v1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*, International Geophysics Series, Vol. 100. 3rd ed., Academic Press, https://doi.org/10.1016/C2009-0-30429-7.

Yu, H., J. Huang, and J. Chou, 2014: Improvement of medium-range forecasts using the analog-dynamical method. *Monthly Weather Review*, **142 (4)**, 1570 – 1587, https://doi.org/10.1175/MWR-D-13-00250.1, URL https://journals.ametsoc.org/view/journals/mwre/142/4/mwr-d-13-00250.1.xml.

Zeng, Y., T. Janjić, A. de Lozar, C. A. Welzbacher, U. Blahak, and A. Seifert, 2021: Assimilating radar radial wind and reflectivity data in an idealized setup of the cosmo-kenda system. *Atmospheric Research*, **249**, 105 282, https://doi.org/https://doi.org/10.1016/j.atmosres.2020.105282, URL https://www.sciencedirect.com/science/article/pii/S0169809520312199.

Zhang, K., 2022: Detectability of labeled weighted automata over monoids. *Discrete Event Dynamic Systems*, **32**, https://doi.org/10.1007/s10626-022-00362-8.

Zhang, Y., N. Liu, and D. S. Oliver, 2010: Ensemble filter methods with perturbed observations applied to nonlinear problems. *Computational Geosciences*, **14**, 249–261, https://doi.org/10.1007/s10596-009-9149-7, URL https://doi.org/10.1007/s10596-009-9149-7.

Zhou, Y., 2024: The 2021 pacific northwest heat wave: Meteorological interpretation of forecast uncertainties in data-driven and physics-based ensembles. M.S. thesis, Karlsruhe Institute of Technology (KIT), Department Troposphere Research (IMK-TRO).

Zängl, G., D. Reinert, P. Rípodas, and M. Baldauf, 2015: The ICON (ICOsahedral non-hydrostatic) modelling framework of DWD and MPI-m: Description of the non-hydrostatic dynamical core. *Quarterly Journal of the Royal Meteorological Society*, **141 (687)**, 563–579.

Acknowledgments

I would like to sincerely thank my supervisors Prof. Dr. Peter Knippertz, Dr. Julian Quinting, and Dr. Jannik Wilhelm, for giving me the opportunity to work on this project and for their deeply valuable guidance and support. I am truly grateful for being a part of the TEEMLEAP project, as it was a very enriching chance that allowed me to be in insightful discussions with brilliant researchers.

A special thank to Dr. Jannik Wilhelm for his continuous support, always available for any questions. He always encouraged my work, inspiring me to approach the problems with calm and rational thinking. His kindness and dedication have meant a lot to me.

I am also grateful to the IMK-TRO Atmospheric Dynamics group for welcoming me, from the meetings to the activities I took part in. A special thank to Prof. Dr. Peter Knippertz for leading and always caring to ensure that it is a comfortable and active team. I truly appreciated the atmosphere and dynamics of the group. A special thank to Dr. Julian Quinting, too, for his continuous support and for his suggestions, and for being always available and attentive for any question or discussion.

I would also like to thank Dr. Sebastian Lerch for the interesting discussions on my project, as well as Yangfan Zhou, Nina Horat, and Dorothea Meike Schwärzel for their valuable insights.

Finally, but not less important, I am deeply grateful to my family and friends for always being so supportive and warm.

Erklärung

Ich versichere wahrheitsgemäß, die Arbeit selbstständig verfasst, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde sowie die Satzung des KIT zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Karlsruhe, den 12.02.2025

Isabel Pena Sánchez