JAMES | Journal of Advances in Modeling Earth Systems®

RESEARCH ARTICLE
10.1029/2024MS004881

TEEMLEAP—A New Testbed for Exploring Machine Learning in Atmospheric Prediction for Research and Education

J. Wilhelm[1,2], J. Quinting[1], M. Burba[2], S. Hollborn[2], U. Ehret[3], I. Pena Sánchez[1], S. Lerch[4,5], J. Meyer[6], B. Verfürth[7], and P. Knippertz[1]

[1]Institute of Meteorology and Climate Research Troposphere Research (IMKTRO), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, [2]German Meteorological Service (DWD), Offenbach, Germany, [3]Institute for Water and Environment—Hydrology, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, [4]Institute of Statistics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, [5]Department of Mathematics and Computer Science, University of Marburg, Marburg, Germany, [6]Scientific Computing Center (SCC), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, [7]Institute for Numerical Simulation, University of Bonn, Bonn, Germany

Key Points:

- A weather forecasting testbed based on the experimentation system of the German Meteorological Service was set up for research and teaching
- It allows studying the entire forecast chain from observations over assimilation and forward integration to verification and post-processing
- The testbed allows a straightforward integration of new machine learning approaches and a fair comparison with conventional methods

Correspondence to:
P. Knippertz,
peter.knippertz@kit.edu

Citation:
Wilhelm, J., Quinting, J., Burba, M., Hollborn, S., Ehret, U., Pena Sánchez, I., et al. (2025). TEEMLEAP—A new testbed for exploring machine learning in atmospheric prediction for research and education. *Journal of Advances in Modeling Earth Systems*, *17*, e2024MS004881. https://doi.org/10.1029/2024MS004881

Received 19 DEC 2024
Accepted 9 MAY 2025

Author Contributions:
Conceptualization: J. Quinting, U. Ehret, S. Lerch, J. Meyer, B. Verfürth, P. Knippertz
Data curation: J. Wilhelm, J. Quinting, M. Burba, S. Hollborn
Formal analysis: J. Wilhelm
Funding acquisition: J. Quinting, U. Ehret, S. Lerch, J. Meyer, B. Verfürth, P. Knippertz
Investigation: J. Wilhelm, M. Burba, I. Pena Sánchez
Methodology: J. Wilhelm, J. Quinting, M. Burba, S. Hollborn, U. Ehret, S. Lerch, J. Meyer, B. Verfürth, P. Knippertz
Project administration: P. Knippertz
Resources: J. Wilhelm, J. Quinting, J. Meyer

Abstract

In the past 5 years, data-driven prediction models and Machine Learning (ML) techniques have revolutionized weather forecasting. Meteorological services around the world are now developing ML components to enhance (or even replace) their numerical weather prediction systems. This shift creates new challenges and opportunities for universities and research centers, calling for a much closer cooperation of meteorology with mathematics and computer sciences, updates of teaching curricula, and new research infrastructures and strategies. To address these challenges, an interdisciplinary team of scientists from the Karlsruhe Institute of Technology (KIT) and the German Meteorological Service (DWD) created the TEstbed for Exploring Machine LEarning in Atmospheric Prediction (TEEMLEAP). Implemented on KIT's supercomputer HoreKa, the TEEMLEAP testbed simulates the entire operational weather forecasting chain using ERA5 reanalysis data as pseudo-observations and DWD's Basic Cycling environment for conducting assimilation-prediction-cycling experiments. Moreover, first steps are taken toward the integration of new data-driven components like FourCastNet and ML-based post-processing methods. The TEEMLEAP testbed allows systematic investigation of a wide range of issues related to weather forecasting such as optimizing the observational system, uncertainty quantification, and developing hybrid systems that integrate ML with physics-based models. This document outlines the testbed's setup, demonstrates its functionality with a pilot experiment, and discusses examples of potential applications. Future plans include creating educational modules and developing a higher-resolution regional version of the testbed that could be used for assimilating field campaign observations.

Plain Language Summary

In the past 5 years, new weather prediction models and Machine Learning (ML) have substantially changed the way we do weather forecasting. Meteorological services worldwide are increasingly using ML to improve their weather prediction systems. This change presents new challenges and opportunities for meteorologists in universities and research centers, as they need to work more closely with mathematics and computer science, update their teaching curricula and create new research facilities and strategies. To better cope with this new situation, scientists from the Karlsruhe Institute of Technology (KIT) and the German Meteorological Service (DWD) created a testbed to explore new approaches to weather forecasting in a systematic way. The system contains key elements of an operational service but in a simplified, easy-to-handle and -understand set-up. In addition to conventional methods, a new ML model is already implemented. The testbed helps studying important issues like improving observation systems, understanding forecast uncertainties, and creating hybrid systems that combine ML with traditional models. This document explains how the testbed works, shows a pilot experiment, and discusses examples of possible use. Future plans include applying the testbed for education and developing a more detailed regional version of the testbed.

1. Introduction

Weather forecasting is an integral part of our daily life and influences a multitude of economic and societal decisions around the world. The increasing number of extreme events and a greater use of renewable energy

Software: J. Wilhelm, J. Quinting, M. Burba, S. Hollborn
Supervision: J. Quinting, S. Hollborn, U. Ehret, S. Lerch, J. Meyer, B. Verfürth, P. Knippertz
Validation: J. Wilhelm
Visualization: J. Wilhelm
Writing – original draft: J. Wilhelm
Writing – review & editing: J. Wilhelm, J. Quinting, M. Burba, S. Hollborn, U. Ehret, I. Pena Sánchez, S. Lerch, J. Meyer, B. Verfürth, P. Knippertz

will further increase our dependence on reliable weather and environmental forecasts (Bauer et al., 2021). Despite more than half a century of scientific and technical advances leading to gradual improvement of operational systems (Bauer et al., 2015) of numerical weather prediction (NWP), weather forecasts today still suffer from systematic errors and inaccuracies (e.g., Frassoni et al., 2023; Hemri et al., 2014; Palmer, 2019; Pantillon et al., 2017; Quinting & Vitart, 2019; Vogel et al., 2018). As a discipline deeply rooted in physical principles and numerical modeling, meteorologists only recently started to realize the enormous potential that lies in the use of artificial intelligence and machine learning (ML) methods (e.g., Boukabara et al., 2021; Dueben & Bauer, 2018; Haupt et al., 2021; McGovern et al., 2019; Reichstein et al., 2019; Vannitsem et al., 2021; Weyn et al., 2019).

In January 2021, the European Center for Medium-Range Weather Forecasts (ECMWF) published a 10-year roadmap to integrate ML into their operational suite (Dueben et al., 2021), which together with the publication of the WeatherBench data set (Rasp et al., 2020, 2024) stimulated interest and investment in the development of ML models for weather forecasting (e.g., Ben Bouallègue et al., 2024; Bonavita et al., 2023). Recently, data-driven weather forecasting models based on ML methods, mainly developed by major technology companies, have exhibited remarkable advances in their forecasting skills. Since 2022, numerous data-driven models with different architectures have been developed, representing significant advancements in the field (e.g., Bi et al., 2023; K. Chen et al., 2023; L. Chen et al., 2023; Lam et al., 2023; Lang et al., 2024; Lessig et al., 2023; Nguyen et al., 2023; Pathak et al., 2022; Price et al., 2024). In contrast to NWP models based on physical laws governing atmospheric processes, these models are trained on ERA5 reanalysis data (Hersbach et al., 2020), and thus learn spatial and temporal patterns and relationships between variables from these data in a statistical sense. While these models often require significant computational resources during training, trained models deliver extremely fast predictions in inference mode, thus reducing computational and energy costs of running forecasts by several orders of magnitude (de Burgh-Day & Leeuwenburg, 2023). These models focus on weather forecasting tasks from the medium to sub-seasonal range and exhibit forecast scores comparable with (and even better than) deterministic state-of-the-art NWP models (Ben Bouallègue et al., 2024; Bonavita, 2024; Charlton-Perez et al., 2024; Rasp et al., 2024). A further major development is the probabilistic hybrid global circulation model that combines a differentiable solver for atmospheric dynamics with machine-learning physics components (Kochkov et al., 2024). It can generate forecasts of deterministic weather, ensemble weather and even climate on par with the best machine-learning and physics-based methods. All these features make data-driven weather forecasting models a compelling alternative to traditional numerical methods. However, thorough evaluation is essential to ensure that forecasters can properly interpret these models and that public needs can be safely and effectively met (Ebert-Uphoff & Hilburn, 2024; McGovern et al., 2024).

Most of these data-driven models can only be used for the forward-integration of the atmospheric state (forecasts) from a pre-determined initial condition (analysis field) from NWP, while the implementation of assimilation-prediction cycles as in operational NWP systems and the development of new ML-based data-assimilation methods has only just begun (Keller & Potthast, 2024; Xiao et al., 2024; Xu et al., 2024). Very recently, McNally et al. (2024) proposed another approach to ML-based weather forecasting: Unlike the data-driven models mentioned above, they trained a neural network to predict future weather purely from historical observations with no dependence on a physics-based model or reanalysis data set (until now up to +12 hr leadtime only).

The aforementioned studies thus followed ECMWF's visionary and ambitious roadmap at an astonishing speed, and started to scrutinize all key elements in the process chain of modern weather forecasting: (a) Observations including corresponding operators, (b) data assimilation including quality control and bias correction, (c) numerical weather prediction (NWP), (d) statistical post-processing, and (e) verification. They identified challenges in research culture, software, and hardware that can only be overcome by closer collaborations between meteorologists, mathematicians, and computer scientists. While ECMWF, the German Meteorological Service (DWD) and other weather centers naturally concentrate on improvements of operational systems in their full complexity, the shift to ML has fundamental implications for universities and research centers active in atmospheric and climate research. It requires establishing a new interdisciplinary culture in research and education of the next generation of weather and climate scientists, which in turn is of high interest for the personnel recruitment of meteorological services. The meteorological services meanwhile are working closely together to integrate ML applications sustainably and profitably into the entire value chain, so for example, in the newly established EUMETNET Artificial Intelligence Program (WMO, 2024). The collaboration spans the areas of data curation, analysis, modeling, and post-processing, as well as products and services.

The fast and revolutionary changes and the emergence of big technology companies with large teams and IT resources in the field of weather forecasting have created substantial risks for universities to fall behind in international leading meteorological research and infrastructure development, but also open new opportunities due to the short runtime of pre-trained ML-based forecasting models. At the same time, traditional education in meteorology is not designed to teach students the skills they need in this fast changing employment landscape. Already in 2021, scientists at KIT realized the risks and chances inherent in this situation and, as a response tailored to university needs, created the idea of TEEMLEAP, a TEstbed for Exploring Machine LEarning in Atmospheric Prediction to explore ML methods in atmospheric prediction with their strategic partner DWD. Mainly based on state-of-the-art NWP components of DWD, this testbed is simple enough to allow systematic analyses along the entire process chain of weather forecasting, yet complex enough to allow extensions to relevant scales and real-world problems. It has great potential to provide new insights into many fundamental issues such as optimization of the observational system, uncertainty quantification and development of hybrid systems integrating ML components with the existing physics-based models. Already now, it allows, inter alia, studies of the sensitivity regarding different observation characteristics as well as investigations of predictions with the data-driven model FourCastNet based on initial conditions obtained from the NWP assimilation cycle. Through addressing highly topical research questions, the testbed can serve as a catalyst for interdisciplinary dialog and collaboration between meteorologists, mathematicians, and computer scientists. This should foster a better understanding of disciplinary cultures and languages, and help identify synergies and necessary theoretical, methodological, and educational development.

The main goal of this paper is to describe the idea behind the testbed, how this idea was transformed to an actual infrastructure at KIT, and how it is being used to address exciting research questions. Testbed components that build upon existing code and models will be characterized concisely, new developments (especially regarding the observation part of the forecast chain) will be presented in more detail.

The remainder of this article is organized as follows: Section 2 describes the data and models, upon which the testbed is built. Its structure and implementation, general settings and simplifications made are presented in Section 3. First applications with the testbed are showcased in Section 4. The paper ends with a summary and conclusions in Section 5.

2. Data and Models

2.1. ERA5 Reanalysis Data

The ECMWF provides the widely used reanalysis data set ERA5, which represents an estimate of past atmospheric states (Hersbach et al., 2020). The native vertical discretization is realized on 137 hybrid sigma model levels between the surface and approximately 100 km altitude. Vertical grid spacing increases from around 50 m near the surface to around 500 m in the lower stratosphere at 20 km altitude. Data comprise key meteorological variables such as temperature, wind, and specific humidity at a native resolution of TL639 (31 km) from 1979 to present. For the purpose of this study all ERA5 data were retrieved on a regular latitude-longitude grid of $0.25° \times 0.25°$ from 27 August to 07 October 2022. Moreover, a multi-year archive for the whole period since 1979 is available at KIT on a regular latitude-longitude grid of $0.5° \times 0.5°$.

2.2. DWD's Basic Cycling Environment BACY

At DWD, an extensive data assimilation and forecasting basic cycling environment (BACY) has been developed for experimentation purposes and applied for various assimilation-related studies (e.g., Bick et al., 2016; Schraff et al., 2016; Zeng et al., 2021). It mimics the operational workflow by organizing different tasks along the NWP forecast chain: data assimilation, forward integration (the actual NWP), and verification. It supports global and regional (limited-area mode) modeling for the atmosphere, including also ocean modeling and climate projections, and incorporates several data assimilation methods (cf. Section 2.2.1). Due to its modular design, it can be easily adapted to other high-performance computing (HPC) environments and specific applications.

The tasks are carried out in separate cycles, from which the key elements for global deterministic experiments are:

- Assimilation cycle (ASM): repeated execution of actual assimilation with the Data Assimilation Coding Environment (DACE; Section 2.2.1) and short-range first-guess forecasts with the Icosahedral Nonhydrostatic modeling framework (ICON; Section 2.2.2). Surface analyses (snow, soil moisture and sea surface temperature) are calculated separately;
- Forecast cycle (MAIN): actual medium-range NWP with ICON, based on initial conditions from the ASM cycle;
- Verification cycle (VERI): calculation of model equivalents of forecasts to be compared with meteorological observations. This comprises spatial and temporal interpolation to in situ observations like radiosoundings, SYNOP and aircraft measurements as well as applying complex observation operators for remote-sensing observations.

As Schraff et al. (2016) pointed out, the BACY environment allows for easy and fast experimentation without accessing a tape-based archiving system, while being fully portable to any standard LINUX-based computing environment. However, observations and model start data for experiments have to be retrieved from DWD's operational archiving system in advance, which is handled by a BACY utility on the DWD HPC infrastructure. For this paper, BACY 1.0 is used (GitLab commit hash 6450a52ce).

2.2.1. Data Assimilation Coding Environment DACE

The DWD has developed DACE as a modular coding environment during the past two decades. It comprises several data assimilation techniques, including variational (3DVar), Particle and Kalman Filter as well as hybrid methods (e.g., the hybrid variational ensemble Kalman filter EnVar). DACE can handle a large variety of state-of-the-art observational data sources, including inter alia SYNOP stations, radiosoundings, aircraft, and remote sensing instruments. In addition, an ample collection of tools for data handling and evaluation, as well as the model equivalent calculator are available. The DACE assimilation method applied in this study (commit hash 7ea048ea3; Section 3.3) needs suitable observations of meteorological variables, as well as their associated error estimates, as input for the optimization.

2.2.2. Icosahedral Nonhydrostatic (ICON) Model

ICON has been developed by DWD and the Max Planck Institute for Meteorology and is the operational NWP system of DWD. ICON uses an icosahedral-triangular Arakawa-C grid with one-way and two-way grid-nesting capability, and can be run in global as well as limited-area mode. The prognostic variables are the horizontal velocity component normal to the triangle edges v_n, the vertical wind component w, density ρ, and virtual potential temperature θ_v. Time integration is performed with a two-time-level predictor-corrector scheme that is fully explicit along the horizontal axes and implicit for the terms describing vertical sound wave propagation; see Zängl et al. (2015) for a detailed description of the non-hydrostatic dynamical core. DWD uses the ICON model and the integrated two-way nesting capability operationally on a global scale on 120 vertical levels with a horizontal resolution of 13 and 6.5 km in the refined nest over the European region. The resolution of the ensemble is 26 km/13 km. The regional model covers an area in Central Europe with a 2 km resolution. BACY experiments can be conducted with and without nesting. Furthermore, it comes with an optimized default setup for ICON close to operations, which contributes to the ease of application for assimilation studies. For this first version of the TEEMLEAP testbed, the ICON release 2.6.6 is used.

2.3. Fourier Forecasting Neural Network (FourCastNet)

The Fourier ForeCasting Neural Network (FourCastNet), developed by Pathak et al. (2022), is a deep-learning, purely data-driven weather forecasting model based on the Vision Transformer architecture with an Adaptive Fourier Neural Operator. In a recent update, Bonev et al. (2023) presented FourCastNet (Version 2) as a further development of the original FourCastNet model. In contrast to the original model, where a flat Euclidian space is assumed, FourCastNet (Version 2) uses Spherical Harmonics Neural Operators for modeling non-linear chaotic and dynamical systems on a sphere with desirable properties such as translational and rotational equivariance. A concise overview is given by Charlton-Perez et al. (2024), and comparisons with other data-driven models (in terms of architecture and forecast evaluation) are, for example, presented in Bonavita (2024), Bülte et al. (2025),

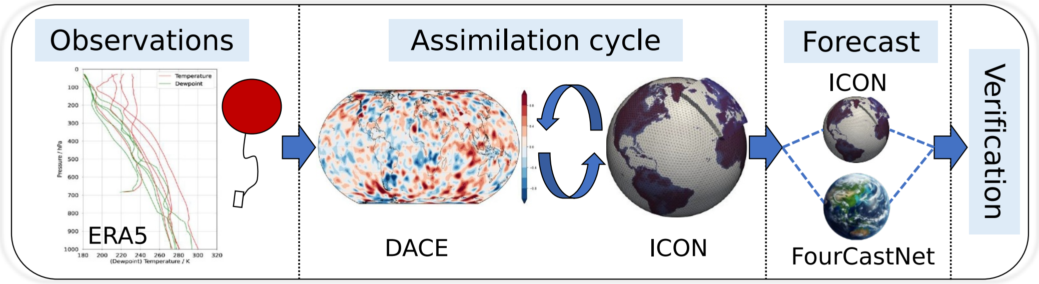

Figure 1. Overview of the TEEMLEAP testbed structure. Vertical profiles from ERA5 reanalysis data mimic radiosounding measurements, which are prepared for the assimilation cycle of DWD's basic cycling environment BACY, where they are assimilated with DACE to create the atmospheric analysis. The ICON model is used for calculating both the assimilation background (first guess) in the assimilation (ASM) cycle and medium-range weather forecasts in the MAIN cycle. Initialized with adapted analyses from the ASM cycle, medium-range weather forecasts can also be calculated with FourCastNet (Version 2). Verification is done on the full ERA5 grid (ICON + FourCastNet), or optionally in observation space by running the VERI cycle (ICON only). The figure of the earth representing ICON is taken from DWD (2024), the earth representing FourCastNet was generated by Microsoft Copilot on 19 September 2024 with the prompt "Earth on a white background."

Feldmann et al. (2024) and Pasche et al. (2024). FourCastNet is trained on several decades of ERA5 data (0.25° horizontal grid-spacing) in a two-step process, consisting of a single autoregressive step followed by fine-tuning. The model is trained with 73 parameters taken from ERA5. These parameters comprise single-level (wind, pressure and temperature close to the surface, mean sea-level pressure, wind in 100 m height), vertically integrated (total column water vapor) and pressure-level variables (wind, geopotential, temperature, relative humidity on 13 levels between 1,000 and 50 hPa). The pre-trained version of FourCastNet (Version 2) used for this study is freely available via ECMWF's *ai-models* (commit hash f51a2d0) and *ai-models-fourcastnetv2* (commit hash a5e2bb2) python packages (cf. Open Research statement).

3. Testbed Structure

3.1. Overview

The heart of the TEEMLEAP weather forecasting testbed is DWD's basic cycling environment BACY (Figure 1; Section 2.2). It provides an environment for assimilation-prediction workflows in a quasi-operational setting with the data assimilation coding envionment DACE (Section 2.2.1) and the numerical weather prediction model ICON (Section 2.2.2). The portability of the environment facilitated its installation on KIT's supercomputer HoreKa (Hochleistungsrechner Karlsruhe; high-performance computer Karlsruhe; https://www.nhr.kit.edu/userdocs/horeka/), a hybrid system with nearly 60,000 processor cores, nearly 300 terabytes of main memory and more than 750 GPUs, with manageable effort. The surface analysis procedures for snow, soil moisture and sea surface temperature (Section 2.2) are technically unrelated to each other and to the atmospheric analysis (DACE), and their implementation on other computers is not straight-forward. Moreover, DWD is currently further developing and unifying the surface analysis procedures. Thus, they are omitted in the first version of the testbed presented in this study. We plan to integrate the unified analyses in the next update of the testbed. Note that cycling experiments without surface analyses can only be run for a limited simulation period, as the missing land-/sea-atmosphere adjustments increasingly affect the realism of the atmospheric state. However, we found still very reasonable results in global 1-month experiments when cycling in the transition period from summer to autumn when, for example, snow cover and sea surface temperature vary only little. The availability of large amounts of memory and dedicated large-memory nodes enables even the assimilation of a high number of observations with DACE, whereas the huge number of processor cores and availability of message-passing interface parallelization is beneficial for time-efficient predictions with the ICON model. The accessibility of GPUs, moreover, was key for the integration of ECMWF's *ai-models* and *ai-models-fourcastnetv2* environments for the application of the data-driven weather prediction model FourCastNet (Version 2) as an additional forecasting option alternative to ICON, yet without the possibility of cycling (cf., Section 1).

In contrast to cycling experiments usually conducted at weather services, the testbed does not make use of real observational data from in situ measurements or remote sensing instruments. These data typically feature spatial inhomogeneities and limited temporal availability, require very complex quality control procedures, and often are subject to legal usage constraints. Instead, the testbed takes advantage of reanalysis data, which serves as a basis for the generation of realistic pseudo-observations. The idea of incorporating simulated observational data stems from Observing System Simulation Experiments performed to study possible improvements of existing and envisioned observing systems (e.g., Andersson & Masutani, 2010; Arnold & Dey, 1986; Errico et al., 2013; Hoffman & Atlas, 2016; Masutani et al., 2010; Privé et al., 2013; Privé et al., 2023). In the first testbed version, the exclusive use of pseudo-radiosoundings allows for easy data handling and a high degree of flexibility to design a wide range of sensitivity experiments. This comes, however, at the price of including observational data only indirectly via the reanalyses, leaving room for model-generated errors and biases. In addition, the restriction to a single observation type reduces the transferability of quantitative results to real-world operational systems, where a variety of observation types with different measured variables, coverage, and availability is assimilated simultaneously, and where complex observation operators, bias corrections and data thinning algorithms are used to prepare measurements for the assimilation. Another challenge is that the error profiles associated with the pseudo-radiosoundings are unknown a priori and have to be properly constructed (cf. Section 3.2.2). It is worth pointing out, however, that the errors of real observations are also not known to full extent, as in addition to instrumental error they include contributions of representativeness.

3.2. Generation of Pseudo-Radiosoundings From ERA5 Reanalyses

3.2.1. Pseudo-Observation Profiles

Typically, real radiosounding observation data, ready for operational assimilation at DWD, consist of vertical profiles of temperature T, relative humidity r, horizontal wind (zonal component u and meridional component v), and associated pressure p. Humidity measurements above 275 hPa often raise various quality issues and are only assimilated for few trustworthy sensors at DWD. Note that through the hydrostatic equation the vertical geopotential height (Z) profile measured by radiosoundings contains redundant information and is thus operationally omitted in the assimilation. Only the geopotential height at the surface, Z_{sfc}, is taken into account, mainly for the reason of an improved surface pressure analysis. These are the standard variables available for assimilation of real radiosounding observations.

In the testbed, the same standard variables as in the operational setup are used for assimilation (T, r, u, v and Z_{sfc} at pressure p). Instead of taking real observations, pseudo-radiosounding measurements of meteorological variables are calculated from the ERA5 reanalysis data in a separate python environment independently from DACE and ICON. For real radiosoundings, vertical resolution is quite high with a measurement interval of 1 s and mean ascent speeds of around 5 m s^{-1}. In the testbed, vertical profiles are created from ERA5 data on the hybrid model levels from near the Earth's surface (level 137) to approximately 24 km altitude (level 40). From the required standard variables for the pseudo-radiosounding measurements, only T, u, v and specific humidity q are archived together with surface pressure and coefficients defining the model levels, from which p, Z_{sfc} (here: Z on the lowermost ERA5 model level 137) and r are derived. The ERA5 values are horizontally interpolated to the predefined latitude-longitude positions of the pseudo-radiosounding stations using a distance-weighted procedure by default. Generally, arbitrary spatial distributions of the pseudo-radiosounding locations can be easily implemented. The default for global sensitivity experiments is the spherical distribution on a constructed Fibonacci lattice (e.g., González, 2010; Swinbank & Purser, 2006), which distributes the pseudo-radiosoundings almost uniformly and isotropically over the globe (Figure 2). The properties for the construction of pseudo-radiosounding observations are summarized in Table 1.

3.2.2. Pseudo-Observation Error Profiles

A big challenge but a necessary step for establishing a functioning cycling framework for the TEEMLEAP testbed is to define adequate observation errors for the pseudo-observations. Standard observation error profiles $\sigma_{ERA5}(p)$ have to be connected to the observation profiles prior to assimilation. These error profiles are essential for the optimization procedure in the data assimilation step (cf. Section 3.3). For real radiosoundings, fixed standard error profiles are usually assumed and taken from look-up tables created from past assimilation experiments, and no

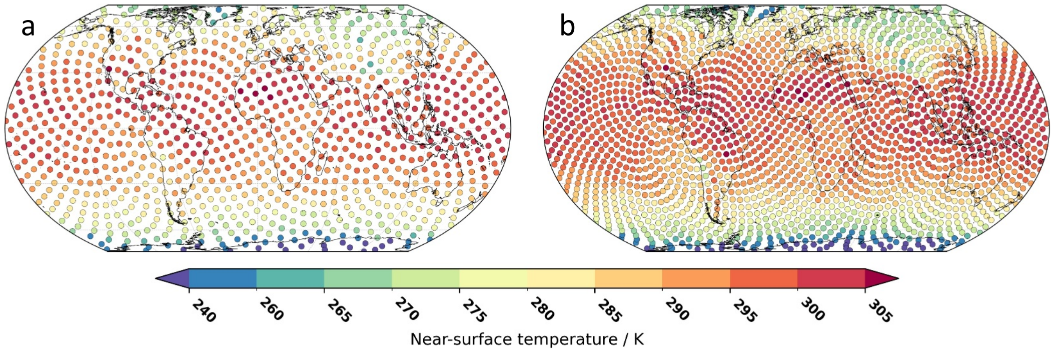

Figure 2. Illustration of the horizontal locations (filled dots) for the pseudo-radiosoundings on a constructed Fibonacci lattice (a) for 1,000 stations and (b) for 2,000 stations. The dots are colored according to the near-surface temperature (lowest ERA5 model level) on 30 September 2022 (00 UTC). The construction yields an almost uniform and isotropic distribution and is thus a good approximation for equidistantly spaced locations. Note that the construction "poles" can be at arbitrary geographical positions (here: 90°W and 90°E at the equator)—rotation of the lattice generally results in different locations, but the same areal coverage.

spatial or inter-variable error correlations are assumed. In the testbed, errors for the pseudo-radiosoundings were not known a priori. Adopting the methodology for real radiosoundings, we diagnosed the typical ERA5-intrinsic error profiles by means of the method of Desroziers et al. (2005) for several test cycling periods up to 1 month with $\mathcal{O}(1000)$ pseudo-radiosoundings in an iterative procedure. The intrinsic error profiles are diagnosed on 14 significant pressure levels between 1,000 and 30 hPa and linearly interpolated in $\ln(p)$ onto $K = 1{,}030$ levels between 1,030 and 1 hPa spaced by 1 hPa afterward to obtain $\sigma_{ERA5}(p)$. Although the profiles comprise these K discrete levels only and are no continuous functions, we write (p) to indicate the pressure-dependencies in this section for the sake of clarity. The nearly cost-free Desroziers 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 diagnosed standard error profile statistics are then assumed to represent the standard deviation of the distribution of the individual ERA5-intrinsic error profiles for each meteorological variable, which here is the best-possible (smallest) standard deviation, as the true atmospheric state is not exactly known.

For experiments with assumed lower observational quality, that is, with higher observation errors, these diagnosed standard deviations are multiplied by a pressure-dependent factor $f(p)$, leading to a total standard deviation

$$\sigma(p_k) = f(p_k)\ \sigma_{ERA5}(p_k) \tag{1}$$

Table 1
Collection of Pseudo-Radiosounding Properties in the TEEMLEAP Testbed

Pseudo-radiosounding property	Value
ERA5 resolution	0.25° or 0.5°
Meteorological variables	T, r, u, v, Z_{sfc}*
Horizontal locations	Fibonacci lattice (quasi-uniform) or lat-lon grid
Number of pseudo-radiosoundings	0 to $\mathcal{O}(10000)$*
Vertical locations	Pressure at ERA5 model-levels
Total number of vertical levels	1 to 98*
Observation errors	Flexible through observation perturbations

Note. In general, the total observation number (horizontal locations × vertical levels × variables; marked with a *), as well as the available physical memory on the computing nodes, determine the upper limit for the assimilation. Technically, it also depends on how the user constrains the convergence of the steepest-gradient solver for the minimization problem.

for each pressure level p_k ($k = 1, \ldots, K$), separately for each meteorological variable. For the geopotential height $Z_{sfc}(p)$, the procedure for determining $\sigma_{ERA5}(p)$ remains the same as for T, r and u/v. Moreover, note that we omit an index indicating the meteorological variable in the equations. Simultaneously, the ERA5-derived observation values are perturbed such that their total error statistics (including both intrinsic ERA5 error and perturbation error) match the desired total standard deviation to be assumed in the assimilation. This can be achieved by adding perturbation profiles $\pi(\tilde{p})$ to the observation profiles $o(\tilde{p})$, where $\tilde{p}$ indicates all pressure levels (up to 98; indicated by $\tilde{K}$) where observations are available, which are different for every pseudo-radiosounding. For this purpose, perturbations $\pi(p)$ are first calculated on the K pressure levels between 1,030 and 1 hPa with a resolution of 1 hPa, prescribing zero mean and standard deviations

$$\sigma_{pert}(p_k) = \sqrt{f(p_k)^2 - 1}\;\sigma_{ERA5}(p_k) \tag{2}$$

under the assumption of Gaussian perturbations and error distributions, for which holds:

$$\sigma(p_k)^2 = \sigma_{ERA5}(p_k)^2 + \sigma_{pert}(p_k)^2. \tag{3}$$

Generally, the perturbation profiles $\pi(p)$ can be constructed arbitrarily under the above constraints. However, to ensure plausible and smooth profiles, the default method applies perturbations that are constructed following Houtekamer (1993) and Houtekamer et al. (1996), based on the eigenvectors of a vertical covariance matrix. This covariance matrix $\mathbf{C}$ can be constructed from the respective perturbation standard deviation profile and a $K \times K$ correlation matrix $\mathbf{R}$:

$$\mathbf{C} = \left[\sigma_{pert}(p)\;\sigma_{pert}^{T}(p)\right] \circ \mathbf{R}, \tag{4}$$

where $\circ$ is the elementwise matrix product (Hadamard product). Analogously to Errico et al. (2013), the elements of $\mathbf{R}$ are expressed by a Gaussian-shaped correlation function:

$$R(p_k, p_l) = \exp\left[-\frac{1}{2}\left(\frac{R_d T_0 (\ln p_k - \ln p_l)}{g_0 L_v}\right)^2\right], \tag{5}$$

which implies $R(p_k, p_k) = 1$. Therein, $R_d = 287.05$ J kg^{-1} K^{-1} is the gas constant of dry air, $T_0 = 270$ K an approximate tropospheric mean temperature, $g_0 = 9.80665$ m s^{-2} the gravitational acceleration, and L_v a characteristic vertical correlation length scale. As Errico et al. (2013), we set the defaults to $L_v = 500$ m for temperature T and wind components u/v, and $L_v = 180$ m for relative humidity r, as this quantity tends to have smaller-scale variability in the troposphere due to cloud, precipitation and mixing processes.

Eventually, as Houtekamer et al. (1996) describe, the perturbation profile for each variable and pseudo-radiosounding is calculated by weighting each eigenmode contribution:

$$\pi(p) = \sum_{k'=1}^{K} w_{k'}\,\lambda_{k'}\,v_{k'}(p). \tag{6}$$

Herein, $\lambda_{k'}^2$ and $v_{k'}(p)$ are the k'-th eigenvalue and eigenvector of $\mathbf{C}$, respectively (we write k' here, as we do not iterate over pressure levels but eigendimensions). The weights $w_{k'}$ are randomly drawn from a standard Gaussian distribution with zero mean and unit variance, independently for each variable and pseudo-radiosounding (Houtekamer, 1993). Thus, the input for assimilation are the perturbed observation profiles $o'(\tilde{p}) = o(\tilde{p}) + \pi(\tilde{p})$, and the corresponding (assumed) observation error standard deviation profiles $\sigma(\tilde{p})$, where the values for $\pi(\tilde{p})$ and $\sigma(\tilde{p})$ correspond to the values of $\pi(p)$ and $\sigma(p)$ on the pressure levels closest to the $\tilde{K}$ observed pressure levels (see Figure 3 for an illustration of the perturbation and error profiles).

The procedure for the perturbation generation was tested extensively. Moreover, the validity of the procedure was checked and confirmed by means of 5-day-long test cycling experiments and evaluation of the relative Desroziers

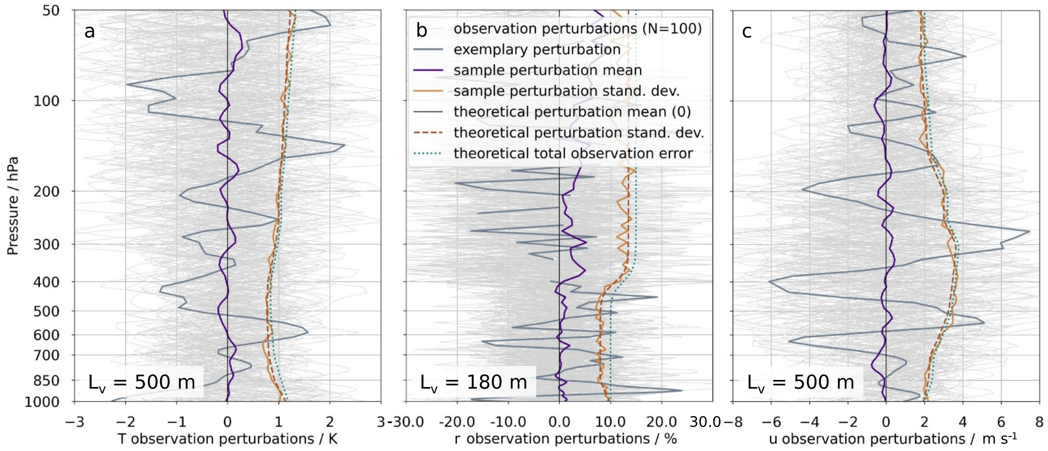

Figure 3. Illustration of exemplary observation perturbations, for (a) temperature T, (b) relative humidity r and (c) zonal wind u. Gray lines depict 100 individual perturbation profiles $\pi(\tilde{p})$ for the respective variables, which are attributed to 100 radiosounding stations. One gray line is drawn thick to help better understand the vertical structure of the profiles. Sample mean and standard deviation of the 100 profiles are drawn in purple and ochre solid lines, respectively. The theoretical mean (0) and standard deviation for the perturbations ($\sigma_{pert}(\tilde{p})$) are drawn as solid black and dashed ochre lines, respectively. The turquoise dotted line represents the theoretical total observation error profiles $\sigma(\tilde{p})$. Theoretical and sample-derived perturbation mean and standard-deviation profiles, as well as the total observation error profile, are obtained by attributing the individual pressure-level values of each sounding to the closest of the $\tilde{K} = 98$ pressure levels corresponding to the ERA5 model level values for a surface pressure of 1,013.25 hPa. The vertical decorrelation lengths L_v used in Equation 5 are given in the bottom-left corner of each sub-plot. In this example, perturbations are calculated such that the total observation error profiles correspond to those operationally used in DACE for real radiosoundings. Missing values in perturbation profiles of r are set, if the perturbed r value would be smaller than 0% or larger than 105% (to balance the allowance for possible supersaturation and the number of missing values).

statistics (not shown). The assumption of Gaussian-distributed errors seems reasonable for all variables except r, which itself is usually not Gaussian-distributed and left- and right-bounded. Improving initial conditions of humidity is a challenge, which is especially important for predicting tropical rainfall patterns more accurately. This challenge could be tackled by applying more costly data assimilation algorithms accounting for non-Gaussian error statistics (e.g., Janjic et al., 2014, 2021; Janjic & Zeng, 2021). In the testbed, however, we currently accept this non-optimal use of humidity observations. For now, after the perturbation profile generation clearly unphysical perturbed r values smaller than 0% or larger than 105% are eliminated. The latter limit balances the allowance for possible supersaturation and an acceptable number of missing r values. Overall, the elimination leads to a marginally positive sample mean for the r perturbations in the upper troposphere of 1%–4% (recall that r is only assimilated below the 275-hPa level in the testbed anyway for now; Figure 3). In the experiments, however, this does not introduce a positive r bias throughout the troposphere even after weeks of cycling, as the assimilation itself acts as a damping mechanism in this regard (not shown).

3.3. Experiment Conduction

In this first version of the testbed, only global experiments are possible. Current and future developments will enable regional experiments as well, but are not discussed further here. Testbed experiments can be of various types: assimilation-only, forecast-only or verification-only experiments; full cycling experiments comprising all components from the generation of pseudo-observations to verification; or experiments with combinations of several components. While BACY cycles can be generally run in parallel, the testbed components are typically carried out sequentially in full cycling experiments for now. Thus, for the desired cycling period, all pseudo-observations are calculated before the ASM cycle begins. The MAIN cycle awaits the completion of the ASM cycle before its start. Likewise, the VERI cycle awaits the completion of MAIN.

Full cycling experiments can be preceded by a spin-up period of customizable length, in which best-possible pseudo-radiosounding data (cf. Section 3.2) are assimilated. This aims to force the ICON background state

Table 2
Overview of the Naming for the Four 7-Day Forecasts With the Physics-Based ICON and the Data-Driven FourCastNet Weather Prediction Models, as Used for the Pilot Experiment Based on Assimilating Data From 1,000 or 2,000 Pseudo-Radiosoundings, Respectively

	Forecasting model	
Observation number	ICON	FourCastNet
1,000	ICON_1000	FCN_1000
2,000	ICON_2000	FCN_2000

toward the ERA5 state. During both the spin-up and the actual experiment period, the 3DVar-based physical-space assimilation system is used for the analysis calculation. The assimilation window and frequency are fixed to 3 hr, as the background error covariances are only available for 3-hr ICON backgrounds. For the sake of smooth state developments, the analysis increments are added using ICON's incremental analysis update within a symmetric 3-hr time window. Medium-range ICON forecasts and verification tasks can be conducted in arbitrary frequency (in multiples of 3 hr). ICON is run in a global default setup used at DWD for NWP-related cycling experiments (without nesting and adaptive parameter tuning, but with full physics), which is close to the operational setup. Setup details are included in the BACY version used (see Section 2.2). Diverse output processing tools for quick looks and evaluations of data can be run after the BACY-related components finished.

The possible settings of the components, as well as HPC-related instructions for the resource allocation, can be handled by parameters set in one single testbed configuration file. The actual creation of an experiment and the conduction of the different testbed components is managed by a single shell script, which only needs the configuration file as input. This allows for an easy handling and usage of the testbed, which enables an introduction for students in a reasonable amount of time. All necessary code is stored in GitLab repositories. ICON is open-source, and BACY-related components of the testbed, are openly available for research purposes (cf. Open Research Statement below). On KIT's supercomputer HoreKa, the binaries and external data needed for the experiment conduction are accessible via shared directories. Thus, installing the entire testbed machinery with corresponding environments on a new HoreKa account (e.g., for students or partners), only requires a few steps, such that the technical infrastructure for new users is quickly ready-to-go. KIT and DWD staff jointly ensure continuous updates of all components and external data in the framework of an institutional cooperation agreement between the two organizations.

4. Illustration of the Concept

The basic idea behind the testbed is to allow systematic analyses along the entire process chain of weather forecasting. Despite simplifications regarding the observation data described in Section 3, testbed experiments can tackle real-world-problems. One possible application is the investigation of sensitivities of the forecast skill to different observation characteristics (see Section 5 for a detailed discussion). Moreover, as ICON analyses from the ASM cycle of testbed experiments can be easily transformed to initial conditions for FourCastNet predictions with ECMWF's *ai-models* and *ai-models-fourcastnetv2* tool, the type of model (NWP vs. data-driven model) used for the forward integration of medium-range weather forecasts can be chosen freely.

As an illustration of the testbed concept, a pilot experiment is presented in the following. Since the aim of this paper is to demonstrate the set-up of the testbed and its multifunctional applicability for systematic analyses, technical settings are not described down to the smallest detail here. Two ASM cycles are started on 27 August 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. We decided for a recent cycling period in early autumn, where the missing surface analyses might have less impact than in other seasons, as sea-surface temperature and snow cover vary only little (cf. Section 3.1). After a 2-day spin-up period (where ICON is pulled toward ERA5 by assimilating a high number of unperturbed pseudo-radiosoundings), data from 1,000 or 2,000 pseudo-radiosoundings, respectively, available on 98 vertical levels, are assimilated every 3 hr with the physical-space assimilation system. The sounding stations are globally to a good approximation equidistantly distributed (Figure 2) and their observation values are perturbed such that their global observation error statistics agree with typical real-radiosounding statistics. This is realized by level-wise perturbations of the standard observation profiles, using the corresponding error tables for real radiosounding observations from DACE as a reference (Figure 3; see Section 3.2.2). Afterward, a new 3-hr ICON background is simulated. The cycling continues for more than a month and ends on 30 September 2022 (00 UTC). Thus, on this day ICON was informed for more than a month about the estimate of the true atmospheric state only by pseudo-radiosounding data derived from ERA5. After the two ASM cycle simulations finished, one ICON and one FourCastNet medium-range weather prediction were conducted for leadtimes up to 7 days, based on the initial conditions on 30 September 2022 (00 UTC) from the ASM cycles (the four experiment names are given in Table 2). As in the ASM cycles, the 7-day

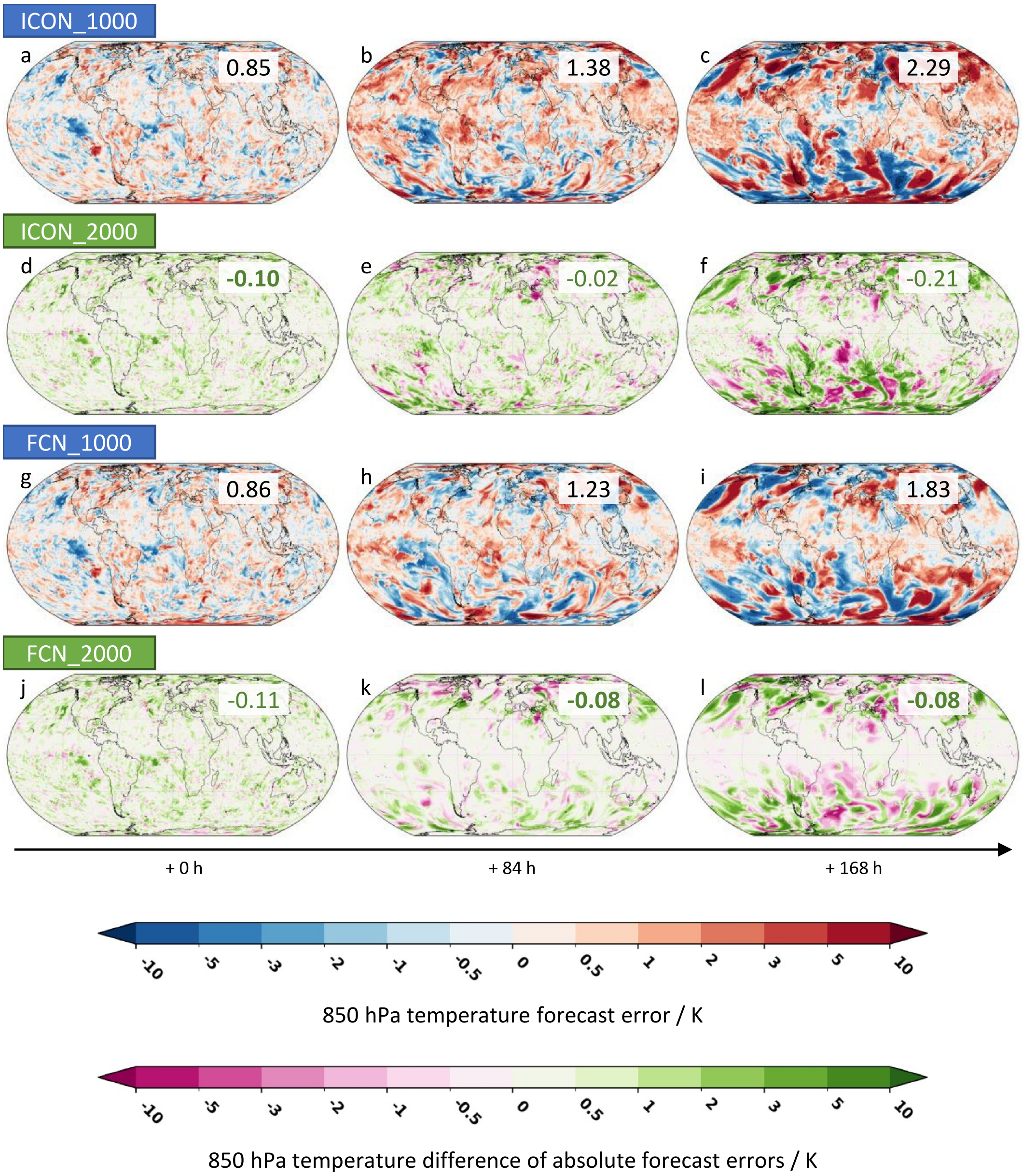

Figure 4. Verification for the pilot experiment of forecasts initialized on 30 September 2022 (00 UTC): 850-hPa temperature forecast error for the ICON_1000 (a–c) and FCN_1000 (g–i) predictions depending on the leadtimes +0 h, +84 h and +168 h (in K). For the ICON_2000 (d–f) and FCN_2000 (j–l) predictions, the respective difference between the absolute values of the forecast errors is displayed, that is, greenish colors indicate a smaller (absolute value of the) forecast error in the simulation with 2,000 pseudo-radiosoundings assimilated compared with the corresponding 1,000-sounding simulation. Numbers give the area-weighted globally averaged absolute value of the forecast error (a–c, g–i) or the corresponding difference as explained above (d–f, j–l). Bold numbers highlight the prediction with lowest forecast error for the respective leadtimes.

ICON forecasts are performed at 26 km horizontal resolution on 90 vertical levels (without nesting). FourCastNet is run in the configuration described in Section 2.3.

We then compare the forecast skill of the four different simulations. Of course, the results cannot be deeply interpreted or generalized—this will be the task of future studies applying systematic analyses of the forecast error over longer time periods. In contrast to operations, where verification is only possible in the space spanned by the available observations, the testbed additionally allows verification on the full ERA5 grid, when the output data of the forecast models is written to the same grid (Figure 4 and Figure A1). In these figures, we highlight the following:

- As expected, the 850-hPa temperature forecasts with ICON exhibit smaller errors (ICON minus ERA5), when data from 2,000 pseudo-radiosoundings are assimilated (Figures 4d–4f), compared with the 1,000-sounding setup (Figures 4a–4c). In this particular case, this is true especially for the longer +168 hr leadtime (on average reduction by 0.21 K or around 9%). Greatest improvements can be seen in mid- and higher latitudes, where baroclinic instability and horizontal temperature advection play a crucial role, as also visible for 500-hPa geopotential height (Figures A1a–A1f).
- Forecasts using FourCastNet have a smaller 850-hPa temperature forecast error than ICON forecasts, although initialized (quasi) from the same analyses (Figures 4g–4l). Note that ICON simulations are initialized applying the incremental analysis update with a 3-hr time window for reasons of a smoother digestion of the analysis increment in the forecasting system to reduce initial noise, which typically results from small-scale non-balanced modes in the analysis (Prill et al., 2024). This explains the tiny differences between ICON and FourCastNet in the +0 hr figures (0.85 vs. 0.86 K error in the 1,000-sounding setup). Interestingly, the error structures in ICON and FourCastNet forecasts are very similar for leadtimes of +84 and +168 hr, emphasizing the realism of the data-driven forecasts. This can also be seen in the 500-hPa geopotential height error patterns (Figures A1g–A1l). Unlike the 850-hPa temperature, globally averaged forecast errors of 500-hPa geopotential height are rather comparable between ICON and FourCastNet simulations in this particular case (e.g., around 17 gpm/41 gpm for +84 hr/+168 hr leadtime in the 1,000-sounding setup).
- Similar to ICON, also the 850-hPa temperature FourCastNet forecasts exhibit smaller forecast errors (FourCastNet minus ERA5), when data from 2,000 pseudo-radiosoundings were assimilated in the preceding ICON-based ASM cycle (Figures 4j–4l), compared with the 1,000-sounding setup (Figures 4g–4i). For +168 hr leadtime, however, the improvement is with 0.08 K smaller than for ICON (but still the globally averaged forecast error is smaller in absolute values: 1.75 vs. 2.08 K).

This exemplary verification can, of course, be extended to further objective forecast error measures as they are typically used on score cards, for instance.

Resource consumption greatly differs between the ICON and FourCastNet forecasts on KIT's supercomputer HoreKa. While one single +168 hr prediction run of ICON (R03B06) needs approximately 16 min of runtime (on 32 nodes with 76 cores) and thus consumes on average 650 CPUh (~5.5 kWh energy consumption), the same FourCastNet prediction runs for around 2 min on a single GPU (~0.01 kWh energy consumption). This means, a FourCastNet forecast only consumes 0.2% of energy compared with ICON, while being one order of magnitude faster. For the experiment, of course, the resources consumed by the ASM cycle have to be considered as well. The DACE assimilation jobs are memory-intensive, but the consumed CPUh differ only slightly between the 1,000- and the 2,000-sounding setup. In total, the DACE and ICON jobs in the 1-month ASM cycle of the pilot experiment consume around 5,000 CPUh (~50 kWh energy consumption) per setup.

5. Summary, Discussion and Outlook

During the past 5 years, the fast progression of data-driven weather prediction models and the increasing use of Machine Learning techniques along the entire weather forecasting chain have fundamentally changed the way how meteorologists will advance the field in the near and far future. These revolutionary changes and the emergence of big technology companies with large teams and IT resources in the field of weather forecasting have created substantial risks for universities to fall behind in international leading meteorological research and infrastructure development but also new opportunities for weather research, as ML-based codes run much faster than conventional methods. At the same time, traditional education in meteorology is not designed to teach students the skills they need in this fast changing employment landscape. As a response tailored to university needs, we implemented the TEEMLEAP testbed on KIT's supercomputer HoreKa to easily perform physics-based cycling experiments, and to explore data-driven models and further ML-based methods in atmospheric prediction with our strategic partner DWD.

This article presented the general set-up of the TEEMLEAP weather forecasting testbed, which mimics the entire operational weather forecasting chain of weather services in a somewhat simplified—yet realistic—manner. Moreover, we illustrated by means of a pilot experiment, how the testbed can be used for tackling highly topical research questions related to weather forecasting in the very dynamic times of the artificial intelligence revolution. The testbed allows systematic investigation of many fundamental issues: How should one design the observational system? How large are the uncertainties in weather prediction and where do they come from? How can we best combine ML components with the existing physics-based models?

We firstly note that with ERA5 reanalyses providing the basis for the generation of pseudo-observations and the verification fields, the possibilities for systematically exploring potential improvements in weather forecasting systems from an academic perspective are manifold. The perturbations of pseudo-radiosounding profiles allow experiments with observations of different quality. Even if the Gaussian assumption for the observation errors simplify the actual problem, Desroziers diagnostics confirm that this assumption gives at least very plausible statistics. While developing the methodology to generate observation error profiles, as described in Section 3.2.2, we recognized that this was a crucial part for the testbed development that needed high scientific diligence and creativity.

Secondly, for long simulation periods, surface analyses are important to account for land-/sea-atmosphere adjustments. These components cannot be easily ported from the DWD operational system, however, and their implementation is planned for the next upgrade of the TEEMLEAP testbed. We found in global 1-month experiments reasonable results when cycling in the transition period from summer to autumn when, e.g., snow cover and sea surface temperature vary comparably little.

The third point worth mentioning is that we constrain ourselves to only one observation type, namely pseudo-radiosoundings, and that we distribute these quasi homogeneously over the globe. This should be seen as a starting point of a longer expedition, which will include more diverse pseudo-observations in the future. However, already for pseudo-radiosoundings, forecast sensitivities can be explored related to, e.g.,:

- The spatial distribution of stations: What if one would increase the station number only over land? How would the forecast skill increase in the tropics, when there were more observations, e.g., in currently data-sparse Africa and the Maritime Continent? How does the NWP model ICON react to the different distributions, and how do data-driven models?
- The accuracy of observations: What if we only had a few high-quality pseudo-radiosoundings as compared to having many low-quality observations? How many profiles need to be assimilated in order to increase the global forecast quality further? What can we expect from new cheap sensor technology or autonomous drone systems?
- The availability of observations in sensitive regions: Numerous methods have been developed to specifically collect observations in sensitive regions and thus reduce uncertainty in the initial atmospheric state (e.g., Langland et al., 1999; Majumdar, 2016). In data denial experiments, the added value of such observations has been shown to depend on the region, season, and observing system (Buizza et al., 2007), and it is not clear what the impact in data-driven models would be. The testbed now allows virtual measurement campaigns to be carried out systematically allowing for data denial experiments in NWP as well as data-driven models.

A study systematically investigating various forecast sensitivities of ICON and FourCastNet including inter alia the number and the quality of observations (different observation error characteristics) as well as ICON model resolution is already under way. Moreover, we are currently investigating the influence of additional pseudo-radiosounding observations in the tropics on African rainfall prediction.

Fourth, with the described observation perturbation method, it is possible to create ensembles of perturbed observations with different perturbation amplitude and/or different observation number. This opens up spaces for the generation of ensembles of smooth and balanced initial conditions as input for deterministic data-driven models, potentially with more advantages than perturbing the initial conditions themselves, independent of the underlying data assimilation procedure. This will help creating expedient probabilistic forecasts with deterministic data-driven models and quantifying the corresponding forecast uncertainties. A first set of tests on the development of such an ensemble-generation procedure for the testbed has just been completed. Another extension of the testbed could be the integration of new data assimilation methods or new data-driven assimilation-prediction systems. Since within BACY other assimilation techniques, for example, a localized ensemble transform Kalman filter, and ICON in limited-area mode are available, future experiments can also focus on regional forecast skills, potentially even assimilating field campaign observations. Moreover, we will further work on integrating statistical and ML-based post-processing methods in the testbed.

Fifth, the testbed has the potential to serve as a supporting infrastructure for testing data-driven models trained by KIT researchers and (future) project partners. Through connecting people and projects, it already now strengthens the ties between KIT and operational weather services. Its methodology can even be extended to simulate the weather in future climate conditions with the NWP model ICON or data-driven models, when using climate projection data as a basis for the generation of pseudo-observations.

Sixth, running the TEEMLEAP testbed was possible through the availability of HPC resources. We used KIT's supercomputer HoreKa, which is part of the National High Performance Computing Alliance and which can be used by scientists free of charge given a successful application for computing time. If such resources are not available at universities or if allocation procedures are cumbersome, cooperation with weather services might be helpful. Sharing computing resources and exchanging information on possible system improvements can result in attractive win-win situations. One lesson we learned was that it is technically challenging—but manageable—to set up the DWD system on a university computer. In order to keep the testbed running with up-to-date components, however, it is expedient to keep track with new developments at DWD (such as changes in model configuration and new requirements regarding initial and boundary data). For this, regular personal status exchange with the DWD developers is very helpful. Running testbed experiments on an HPC with batch system such as HoreKa requires wise choices of resource settings (runtimes, CPU-hours, nodes, memory), with estimates gained from a series of simulation tests. Without the possibility of node reservations, queuing times can influence the experiment duration significantly, as assimilation and prediction jobs are submitted separately owing to their very different resource requirements.

And last but certainly not least, the testbed is intended to play an important role in educating students from meteorology, mathematics and computer science, as it unites practical applications of an NWP system and of data-driven weather prediction models and ML methods, the relevance of which in weather prediction has dramatically increased during the past few years and will do for the foreseeable future. Two Master's students have already successfully conducted testbed experiments within their projects. In the near future, as part of the meteorological practicals at KIT, entire cohorts of Master's students will conduct own forecast experiments with the TEEMLEAP testbed on the HPC systems at KIT, assimilating weather data they collected with the mobile observation platform KITcube (KIT, 2025).

The manifold research and education opportunities make the testbed a unique, flexible and appealing hands-on experience for future meteorologists from the Master to the senior scientist levels.

Appendix A: 500-hPa Geopotential in the Pilot Experiment

Figure A1.

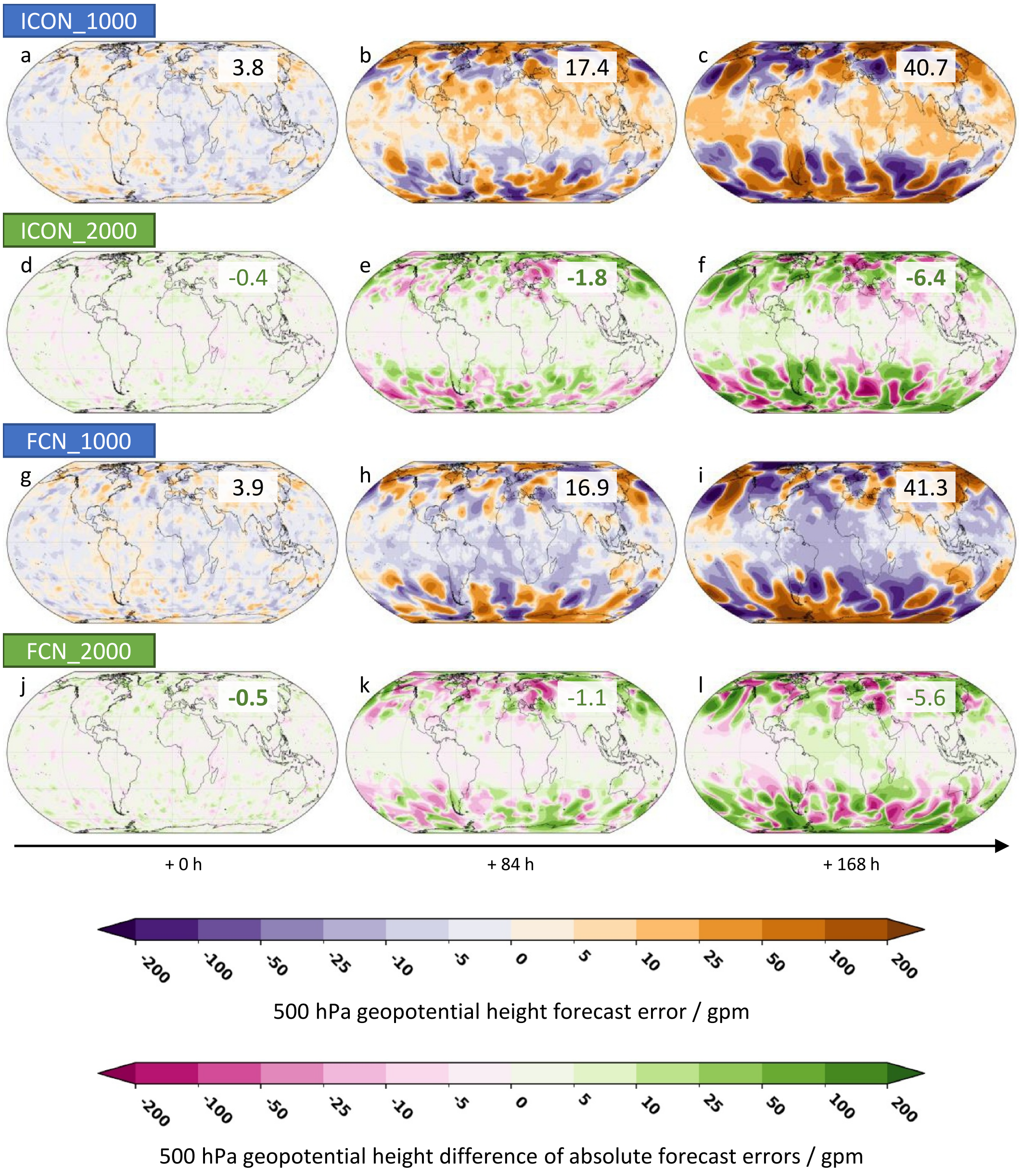

Figure A1. As Figure 4, but for 500-hPa geopotential height (in gpm).

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

ERA5 data are freely available from the Copernicus Climate Data Storage (Hersbach et al., 2017). Code examples of how we conduct experiments with the TEEMLEAP testbed, executable code for the generation of pseudo-radiosoundings and for the calculation of observation perturbations, the Jupyter notebook and data used for the generation of the figures of this manuscript, as well as exemplary data, are openly available at RADAR4KIT (KIT, 2024a; KIT, 2024b). DWD's basic cycling environment BACY as part of DACE is openly available for research purposes, access to the GitLab repositories can be requested from DWD. The ICON model is available under an open-source license (ICON partnership [DWD and MPI-M and DKRZ and KIT and C2SM], 2024). The pre-trained version of the data-driven model FourCastNet (Version 2) used for this study is freely available via ECMWF's *ai-models* and *ai-models-fourcastnetv2* python packages (ECMWF, 2024a; ECMWF, 2024b).

Acknowledgments

This work was realized through the Future Fields (Stage 2) program of the Karlsruhe Institute of Technology (KIT), the Research University in the Helmholtz Association, funded by the Federal Ministry of Education and Research (BMBF) and the Baden-Württemberg Ministry of Science as part of the Excellence Strategy of the German Federal and State Governments. We thank the German Meteorological Service (DWD) for its support, particularly the DWD colleagues Thorsten Steinert, Alexander Cress, Harald Anlauf, Ruud Dirksen, Xu and Michael Krayer. We also acknowledge fruitful discussions with the KIT colleagues Annika Oertel, Nina Horat, Matthias Fischer, Thomas Muschinski, Julia Thomas, Charlotte Debus and Yangfan Zhou. The European Center for Medium-Range Weather Forecasts (ECMWF) and the Copernicus Climate Data Storage (CDS) are acknowledged for providing free access to the ERA5 data set. Moreover, we thank NVIDIA and ECMWF for free provision of the pre-trained data-driven model FourCastNet via the *ai-models* and *ai-models-fourcastnetv2* python packages. This work was performed on the HoreKa supercomputer with the help of the Large Scale Data Facility at KIT, both funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research. Open Access funding enabled and organized by Projekt DEAL.

References

Andersson, E., & Masutani, M. (2010). Collaboration on Observing System Simulation Experiments (joint OSSE). In B. Riddaway (Ed.), *ECMWF newsletter No. 123—spring 2010* (pp. 14–16). ECMWF. Retrieved from https://www.ecmwf.int/sites/default/files/elibrary/042010/14602-newsletter-no123-spring-2010_1.pdf

Arnold, C. P., & Dey, C. H. (1986). Observing-systems simulation experiments: Past, present, and future. *Bulletin of the American Meteorological Society*, *67*(6), 687–695. https://doi.org/10.1175/1520-0477(1986)067<0687:ossepp>2.0.co;2

Bauer, P., Stevens, B., & Hazeleger, W. (2021). A digital twin of Earth for the green transition. *Nature Climate Change*, *11*(2), 80–83. https://doi.org/10.1038/s41558-021-00986-y

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. https://doi.org/10.1038/nature14956

Ben Bouallègue, Z., Clare, M. C. A., Magnusson, L., Gascón, E., Maier-Gerber, M., Janoušek, M., et al. (2024). The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an operational-like context. *Bulletin of the American Meteorological Society*, *105*(6), E864–E883. https://doi.org/10.1175/BAMS-D-23-0162.1

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, *619*(7970), 533–538. https://doi.org/10.1038/s41586-023-06185-3

Bick, T., Simmer, C., Trömel, S., Wapler, K., Hendricks Franssen, H.-J., Stephan, K., et al. (2016). Assimilation of 3D radar reflectivities with an ensemble Kalman filter on the convective scale. *Quarterly Journal of the Royal Meteorological Society*, *142*(696), 1490–1504. https://doi.org/10.1002/qj.2751

Bonavita, M. (2024). On some limitations of current Machine Learning weather prediction models. *Geophysical Research Letters*, *51*(12), e2023GL107377. https://doi.org/10.1029/2023GL107377

Bonavita, M., Schneider, R., Arcucci, R., Chantry, M., Chrust, M., Geer, A., et al. (2023). 2022 ECMWF-ESA workshop report: Current status, progress and opportunities in machine learning for earth system observation and prediction. *npj Climate and Atmospheric Science*, *6*(1), 87. https://doi.org/10.1038/s41612-023-00387-2

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

Boukabara, S.-A., Krasnopolsky, V., Penny, S. G., Stewart, J. Q., McGovern, A., Hall, D., et al. (2021). Outlook for exploiting artificial intelligence in the earth and environmental sciences. *Bulletin of the American Meteorological Society*, *102*(5), E1016–E1032. https://doi.org/10.1175/BAMS-D-20-0031.1

Buizza, R., Cardinali, C., Kelly, G., & Thépaut, J.-N. (2007). The value of observations. II: The value of observations located in singular-vector-based target areas. *Quarterly Journal of the Royal Meteorological Society*, *133*(628), 1817–1832. https://doi.org/10.1002/qj.149

Bülte, C., Horat, N., Quinting, J., & Lerch, S. (2025). Uncertainty quantification for data-driven weather models. *Artificial Intelligence for the Earth Systems*. https://doi.org/10.1175/AIES-D-24-0049.1

Charlton-Perez, A. J., Dacre, H. F., Driscoll, S., Gray, S. L., Harvey, B., Harvey, N. J., et al. (2024). Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of storm Ciarán. *npj Climate and Atmospheric Science*, *7*(1), 93. https://doi.org/10.1038/s41612-024-00638-w

Chen, K., Han, T., Gong, J., Bai, L., Ling, F., Luo, J.-J., et al. (2023a). FengWu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. https://arxiv.org/abs/2304.02948

Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., & Li, H. (2023b). Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, *6*(1), 190. https://doi.org/10.1038/s41612-023-00512-1

de Burgh-Day, C. O., & Leeuwenburg, T. (2023). Machine learning for numerical weather and climate modelling: A review. *Geoscientific Model Development*, *16*(22), 6433–6477. https://doi.org/10.5194/gmd-16-6433-2023

Desroziers, G., Berre, L., Chapnik, B., & Poli, P. (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/10.1256/qj.05.108

Dueben, P., & Bauer, P. (2018). Challenges and design choices for global weather and climate models based on Machine Learning. *Geoscientific Model Development*, *11*(10), 3999–4009. https://doi.org/10.5194/gmd-11-3999-2018

Dueben, P., Modigliani, U., Geer, A., Siemen, S., Pappenberger, F., Bauer, P., et al. (2021). Machine learning at ECMWF: A roadmap for the next 10 years, ECMWF. *Reading, United Kingdom*. https://doi.org/10.21957/ge7ckgm

DWD. (2024). Figure of the icon model grid. Retrieved from https://www.dwd.de/DE/leistungen/nwv_icon_aenderungen/nwv_icon_aenderungen.html

Ebert-Uphoff, I., & Hilburn, K. (2024). The outlook for AI weather prediction. *Nature*, *619*(7970), 473–474. https://doi.org/10.1038/d41586-023-02084-9

ECMWF. (2024a). AI-models. [Code]. Retrieved from https://github.com/ecmwf-lab/ai-models

ECMWF. (2024b). AI-models-fourcastnetv2. [Code]. Retrieved from https://github.com/ecmwf-lab/ai-models-fourcastnetv2

Errico, R. M., Yang, R., Privé, N. C., Tai, K.-S., Todling, R., Sienkiewicz, M. E., & Guo, J. (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/10.1002/qj.2027

Feldmann, M., Beucler, T., Gomez, M., & Martius, O. (2024). Lightning-fast convective outlooks: Predicting severe convective environments with global AI-based weather models. *Geophysical Research Letters*, *51*(22). https://doi.org/10.1029/2024gl110960

Frassoni, A., Reynolds, C., Wedi, N., Bouallègue, Z. B., Caltabiano, A. C. V., Casati, B., et al. (2023). Systematic errors in weather and climate models: Challenges and opportunities in complex coupled modeling systems. *Bulletin of the American Meteorological Society*, *104*(9), E1687–E1693. https://doi.org/10.1175/BAMS-D-23-0102.1

González, A. (2010). Measurement of areas on a sphere using Fibonacci and latitude–longitude lattices. *Mathematical Geosciences*, *42*(1), 49–64. https://doi.org/10.1007/s11004-009-9257-x

Haupt, S. E., Chapman, W., Adams, S. V., Kirkwood, C., Hosking, J. S., Robinson, N. H., et al. (2021). Towards implementing artificial intelligence post-processing in weather and climate: Proposed actions from the Oxford 2019 workshop. *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200091. https://doi.org/10.1098/rsta.2020.0091

Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K., & Haiden, T. (2014). Trends in the predictive performance of raw ensemble weather forecasts. *Geophysical Research Letters*, *41*(24), 9197–9205. https://doi.org/10.1002/2014GL062472

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz Sabater, J., et al. (2017). Complete ERA5 from 1940: Fifth generation of ECMWF atmospheric reanalyses of the global climate [Dataset]. *(Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. https://doi.org/10.24381/cds.143582cf

Hoffman, R. N., & Atlas, R. (2016). Future observing system simulation experiments. *Bulletin of the American Meteorological Society*, *97*(9), 1601–1616. https://doi.org/10.1175/BAMS-D-15-00200.1

Houtekamer, P. L. (1993). Global and local skill forecasts. *Monthly Weather Review*, *121*(6), 1834–1846. https://doi.org/10.1175/1520-0493(1993)121<1834:galsf>2.0.co;2

Houtekamer, P. L., Lefaivre, L., Derome, J., Ritchie, H., & Mitchell, H. L. (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

ICON partnership [DWD and MPI-M and DKRZ and KIT and C2SM]. (2024). Icon release 2024.01. [Code]. *World Data Center for Climate (WDCC) at DKRZ*. https://doi.org/10.35089/WDCC/IconRelease01

Janjic, T., McLaughlin, D., Cohn, S. E., & Verlaan, M. (2014). Conservation of mass and preservation of positivity with ensemble-type Kalman filter algorithms. *Monthly Weather Review*, *142*(2), 755–773. https://doi.org/10.1175/MWR-D-13-00056.1

Janjic, T., Ruckstuhl, Y., & Toint, P. L. (2021). A data assimilation algorithm for predicting rain. *Quarterly Journal of the Royal Meteorological Society*, *147*(736), 1949–1963. https://doi.org/10.1002/qj.4004

Janjic, T., & Zeng, Y. (2021). Weakly constrained Letkf for estimation of hydrometeor variables in convective-scale data assimilation. *Geophysical Research Letters*, *48*(24), e2021GL094962. https://doi.org/10.1029/2021gl094962

Keller, J. D., & Potthast, R. (2024). AI-based data assimilation: Learning the functional of analysis estimation. https://arxiv.org/abs/2406.00390

KIT. (2024a). TEEMLEAP_testbed data – Accompanying data for the first publication on the TEEMLEAP weather forecasting testbed [Dataset]. https://doi.org/10.35097/vdqadxgwx7jjz9w6

KIT. (2024b). TEEMLEAP_testbed_v0.5. [Code]. https://doi.org/10.35097/914u543s5qcnu3nw

KIT. (2025). Kitcube website. Retrieved from https://www.kitcube.kit.edu/english/index.php

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., et al. (2024). Neural general circulation models for weather and climate. *Nature*, *632*(8027), 1060–1066. https://doi.org/10.1038/s41586-024-07744-y

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., et al. (2023). Learning skillful medium-range global weather forecasting. *Science*, *382*(6677), 1416–1421. https://doi.org/10.1126/science.adi2336

Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., et al. (2024). Aifs – ECMWF's data-driven forecasting system. https://arxiv.org/abs/2406.01465

Langland, R. H., Gelaro, R., Rohaly, G. D., & Shapiro, M. A. (1999). Targeted observations in FASTEX: Adjoint-based targeting procedures and data impact experiments in IOP17 and IOP18. *Quarterly Journal of the Royal Meteorological Society*, *125*(561), 3241–3270. https://doi.org/10.1002/qj.49712556107

Lessig, C., Luise, I., Gong, B., Langguth, M., Stadtler, S., & Schultz, M. (2023). AtmoRep: A stochastic model of atmosphere dynamics using large scale representation learning. https://arxiv.org/abs/2308.13280

Majumdar, S. J. (2016). A review of targeted observations. *Bulletin of the American Meteorological Society*, *97*(12), 2287–2303. https://doi.org/10.1175/BAMS-D-14-00259.1

Masutani, M., Woollen, J. S., Lord, S. J., Emmitt, G. D., Kleespies, T. J., Wood, S. A., et al. (2010). Observing system simulation experiments at the national centers for environmental prediction. *Journal of Geophysical Research*, *115*(D7), D07101. https://doi.org/10.1029/2009JD012528

McGovern, A., Bostrom, A., McGraw, M., Chase, R. J., Gagne, D. J., Ebert-Uphoff, I., et al. (2024). Identifying and categorizing bias in AI/ML for Earth sciences. *Bulletin of the American Meteorological Society*, *105*(3), E567–E583. https://doi.org/10.1175/BAMS-D-23-0196.1

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of Machine Learning. *Bulletin of the American Meteorological Society*, *100*(11), 2175–2199. https://doi.org/10.1175/BAMS-D-18-0195.1

McNally, A., Lessig, C., Lean, P., Boucher, E., Alexe, M., Pinnington, E., et al. (2024). Data driven weather forecasts trained and initialised directly from observations. https://arxiv.org/abs/2407.15586

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., & Grover, A. (2023). Climax: A foundation model for weather and climate. https://arxiv.org/abs/2301.10343

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

Pantillon, F., Knippertz, P., & Corsmeier, U. (2017). Revisiting the synoptic-scale predictability of severe European winter storms using ECMWF ensemble reforecasts. *Natural Hazards and Earth System Sciences*, *17*(10), 1795–1810. https://doi.org/10.5194/nhess-17-1795-2017

Pasche, O. C., Wider, J., Zhang, Z., Zscheischler, J., & Engelke, S. (2024). Validating deep-learning weather forecast models on recent high-impact extreme events. https://arxiv.org/abs/2404.17652

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., et al. (2022). Fourcastnet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. https://arxiv.org/abs/2202.11214

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., et al. (2024). Probabilistic weather forecasting with machine learning. *Nature*, *637*(8044), 84–90. https://doi.org/10.1038/s41586-024-08252-9

Prill, F., Reinert, D., Rieger, D., & Zängl, G. (2024). ICON tutorial—working with the ICON model. https://doi.org/10.5676/DWD_pub/nwv/icon_tutorial2024

Privé, N. C., Errico, R. M., & Tai, K.-S. (2013). Validation of the forecast skill of the global modeling and assimilation office observing system simulation experiment. *Quarterly Journal of the Royal Meteorological Society*, *139*(674), 1354–1363. https://doi.org/10.1002/qj.2029

Privé, N. C., McGrath-Spangler, E. L., Carvalho, D., Karpowicz, B. M., & Moradi, I. (2023). Robustness of observing system simulation experiments. *Tellus A: Dynamic Meteorology and Oceanography*, *75*(1), 309–333. https://doi.org/10.16993/tellusa.3254

Quinting, J. F., & Vitart, F. (2019). Representation of synoptic-scale rossby wave packets and blocking in the S2S prediction project database. *Geophysical Research Letters*, *46*(2), 1070–1078. https://doi.org/10.1029/2018GL081381

Rasp, S., Dueben, P., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2020MS002203. https://doi.org/10.1029/2020MS002203

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., et al. (2024). WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, *16*(6), e2023MS004019. https://doi.org/10.1029/2023MS004019

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Schraff, C., Reich, H., Rhodin, A., Schomburg, A., Stephan, K., Periáñez, A., & Potthast, R. (2016). Kilometre-scale Ensemble Data Assimilation for the COSMO model (KENDA). *Quarterly Journal of the Royal Meteorological Society*, *142*(696), 1453–1472. https://doi.org/10.1002/qj.2748

Swinbank, R., & Purser, R. J. (2006). Fibonacci grids: A novel approach to global modelling. *Quarterly Journal of the Royal Meteorological Society*, *132*(619), 1769–1793. https://doi.org/10.1256/qj.05.227

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., et al. (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. https://doi.org/10.1175/BAMS-D-19-0308.1

Vogel, P., Knippertz, P., Fink, A. H., Schlueter, A., & Gneiting, T. (2018). Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa. *Weather and Forecasting*, *33*(2), 369–388. https://doi.org/10.1175/WAF-D-17-0127.1

Weyn, J. A., Durran, D. R., & Caruana, R. (2019). Can machines learn to predict weather? Using deep learning to predict gridded 500-hpa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2680–2693. https://doi.org/10.1029/2019MS001705

WMO. (2024). EUMETNET launches ambitious AI programme with ECMWF and EUMETSAT: WMO research group to enhance collaboration strategies. Retrieved from https://community.wmo.int/en/news/eumetnet-launches-ambitious-ai-programme-ecmwf-and-eumetsat-wmo-research-group-enhance-collaboration-strategies

Xiao, Y., Bai, L., Xue, W., Chen, K., Han, T., & Ouyang, W. (2024). FengWu-4DVar: Coupling the data-driven weather forecasting model with 4D variational assimilation. https://arxiv.org/abs/2312.12455

Xu, X., Sun, X., Han, W., Zhong, X., Chen, L., & Li, H. (2024). FuXi-DA: A generalized deep learning data assimilation framework for assimilating satellite observations. https://arxiv.org/abs/2404.08522

Zängl, G., Reinert, D., Rípodas, P., & Baldauf, M. (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. https://doi.org/10.1002/qj.2378

Zeng, Y., Janjić, T., de Lozar, A., Welzbacher, C. A., Blahak, U., & Seifert, A. (2021). Assimilating radar radial wind and reflectivity data in an idealized setup of the COSMO-KENDA system. *Atmospheric Research*, *249*, 105282. https://doi.org/10.1016/j.atmosres.2020.105282