JGR | Machine Learning and Computation

RESEARCH ARTICLE
10.1029/2025JH001180

Key Points:

- None of the tested AI weather models was able to reproduce the butterfly effect in a physically consistent way
- A first group of AI models showed no sign of accelerated initial perturbation growth and no indication of a fundamental predictability limit
- A second group of AI models did show accelerated initial growth and limited predictability, but this was based on numerical noise only

Correspondence to:
T. Selz,
tobias.selz@kit.edu

Citation:
Selz, T., & Craig, G. C. (2026). Can AI-based weather prediction models simulate the butterfly effect? The role of architecture and implementation. *Journal of Geophysical Research: Machine Learning and Computation*, *3*, e2025JH001180. https://doi.org/10.1029/2025JH001180

Received 9 DEC 2025
Accepted 10 JUN 2026

Can AI-Based Weather Prediction Models Simulate the Butterfly Effect? The Role of Architecture and Implementation

T. Selz[1] **and G. C. Craig**[2]

[1]Karlsruhe Institute for Technology, Karlsruhe, Germany, [2]Ludwig-Maximilians-Universität München, Munich, Germany

Abstract Simulations of numerical weather prediction models indicate that the atmosphere possesses an intrinsic limit of predictability. Initial perturbations of tiny amplitude grow quickly in areas of convection and latent heat release, then spread out and move upscale, eventually affecting even the largest planetary scales after about 2 weeks. In this study, we investigate the ability of several state-of-the-art AI-based weather prediction models to reproduce this phenomenon, which is sometimes referred to as the "butterfly effect." The AI results are compared to those of a conventional, physics-based, weather prediction model run at various resolutions. Evaluating six key characteristics of this butterfly effect, we find that the behavior of the AI models can be separated into two groups. The first group did not reproduce any of the key characteristics, while the second group did reproduce some, in particular fast initial uncertainty growth and indication of an intrinsic limit. However, the behavior was physically inconsistent and based on the production of numerical noise, and for some models even dependent on whether the experiments were carried out on a CPU or a GPU. It seems likely that the inability of AI models to simulate the butterfly effect results from limitations in the analysis data used for training, since their size, design and architecture turned out to be largely irrelevant.

Plain Language Summary This study asks whether today's artificial-intelligence (AI) weather-forecasting models can reproduce the "butterfly effect"—an atmospheric phenomenon where extremely tiny disturbances can grow rapidly, spread across the globe, and ultimately limit how far ahead weather can be predicted. Using several modern AI models, we compare how small perturbations evolve in these systems versus in a traditional physics-based weather model. The physics-based model behaves as expected: Tiny changes grow quickly in regions of storms and heavy rain, spread outward, and eventually affect large-scale weather patterns after about 2 weeks. However, none of the AI models reproduce this physical behavior correctly. Some models show no rapid initial growth at all, behaving as if the atmosphere were more predictable than it really is. Others do show rapid growth, but it is caused by numerical noise, which is an artifact of the computation rather than a real atmospheric process. In some cases, this noise disappears when the models run on a different processor. The results suggest that although AI models can make accurate forecasts, they currently do not capture the true physical limits of predictability governed by the butterfly effect.

1. Introduction

Theoretical arguments and complex numerical simulations of the atmosphere have shown that tiny-amplitude initial perturbations grow extremely fast on small spatial scales, subsequently expanding upscale, and eventually leading to fundamentally limited predictability of the entire atmospheric circulation (e.g., Judt, 2018; Lorenz, 1969b; Selz, 2019). Recently, we showed that the AI weather prediction model "Pangu Weather" was not able to reproduce this behavior and substantially underestimated the uncertainty growth from tiny perturbations (Selz & Craig, 2023). Since our short pilot study on Pangu, many more AI weather models have been released using different architectures, more learned parameters, more training and more computational power. It is therefore interesting to revisit the representation of perturbation growth by AI and how it is affected by these recent developments.

The atmosphere is a chaotic dynamical system, that is, the outcome of a forecast is highly sensitive to uncertainties in the estimation of the initial state. This large sensitivity to the initial conditions, which is also present in low-dimensional systems (e.g., Lorenz, 1963) makes weather forecasting a difficult problem. Nevertheless over the past 50 years, constant improvements of the initial condition estimates and the model have led to longer and better predictions of the weather (Bauer et al., 2015). Unfortunately for weather prediction, the atmosphere is

not a low-dimensional system but composed of many interacting scales of motion which imposes a fundamental limit on predictability beyond the mere sensitivity to initial conditions. Lorenz (1969b) demonstrated by using a simplified turbulence model that uncertainties on small scales can propagate upscale in a cascade-like fashion and so quickly, that even the largest scales become intrinsically unpredictable by about 2 weeks. Following Palmer et al. (2014) we refer to this phenomenon as the "butterfly effect."

Lorenz' work has later been refined based on more realistic simulations from complex numerical models. Zhang et al. (2007) formulated a 3-stage conceptual model, which describes the physical processes that lead to the intrinsic limit. In stage 1, initial uncertainties of tiny amplitude grow extremely fast in regions of convection and latent heat release, leading to a complete decorrelation of the cloud field in only 6–12 hr. In stage 2, these cloud-scale uncertainties spread out by gravity waves and divergent motions and undergo adjustment to geostrophic balance. Finally in stage 3 the geostrophically balanced uncertainties grow further upscale until even planetary scales become decorrelated and all predictability is lost. This conceptual framework does not apply to operational forecasts which are, despite the past progress, still initialized with rather large initial condition uncertainty and immediately trigger stage 3 of the error growth concept. However, near-identical twin experiments confirmed the contribution of convection and moist processes to error growth (stage 1 and 2) and limited intrinsic predictability, once the initial condition uncertainty gets sufficiently small (Baumgart et al., 2018; Judt, 2018; Selz, 2019; Selz & Craig, 2015; Zhang et al., 2019). Selz et al. (2022) estimated that "small" refers to about 10% of the amplitude of current initial uncertainties at which point the upscale growth of uncertainties at small scales (the "butterfly effect") dominates over the direct uncertainty growth on synoptic scales (the mere initial condition sensitivity).

Since the beginning of numerical weather forecasting in the 1950s, models of the atmosphere have relied on numerical solvers of the underlying physically-based partial-differential equations (PDE), with increasing levels of complexity and resolution. Processes that are too small to be resolved or are not part of the fluid equations are addressed with simplified, semi-empirical methods called parameterizations. Recently, an alternative data-driven method for weather prediction has emerged which is based on artificial intelligence and machine learning. However, despite the ability of these new models to produce decent weather forecasts we demonstrated in a small pilot study (Selz & Craig, 2023) that the AI model "Pangu Weather" was unable to reproduce the butterfly effect and instead suggested an unlimited intrinsic atmospheric predictability. This failure was hypothesized to be related to the low effective resolution of Pangu Weather (Selz et al., 2025) or basic properties of the training data set, which prevented rapid perturbation growth on small scales.

Since the release of Pangu, which employs a transformer neural network, the field of AI weather forecasting has evolved quickly. Many more approaches and architectures have been developed and network capacities and resources spent on training have increased. These innovations include models based on graph neural networks (GraphCast; Lam et al., 2023), improved transformers trained on a large amount of data (Aurora; Bodnar et al., 2025), transformers applied to spectral modes (FourCastNet; Kurth et al., 2023) and hybrid approaches, where AI is only used to represent the parameterizations, while fluid motions are computed conventionally (NeuralGCM; Kochkov et al., 2024). In addition, stochastic models have been developed building on denoising techniques, that are also capable of representing forecast uncertainties (GenCast; Price et al., 2025).

In this study, we revisit the representation of the butterfly effect with AI, taking these new developments into account. To do so, we conduct experiments by running ensembles using the AI weather prediction models listed above and in Table 1, initialized from tiny-amplitude initial condition perturbations (as in Selz & Craig, 2023). The AI-based experiments are complemented with simulations from a conventional weather prediction model at various resolutions, ranging from convection-permitting to coarse resolutions which match the smoothness of the AI models. To evaluate the performance of these experiments with respect to the butterfly effect we define a set of six quantities characteristic of the 3-stage error growth concept which are able to distinguish butterfly-like uncertainty growth from uncertainty growth in operational weather forecasts based on much larger uncertainties. We employ additional simulations with varying initial perturbation amplitudes to explicitly test if the predictability time estimated from the different experiments converges toward a finite number, hence if a certain model indicates a fundamental limit to predictability or not.

Table 1
Overview of the Models Used in This Study and Some of Their Basic Parameters

Model	Architecture	Grid size	Time step	Training	n params.	Hardware
ICON	conv. PDE solver	2.5–80 km	4.5–144 s	n/a	n/a	CPU
Pangu	Transformer	0.25°	6 h, 24 h	ERA5	2.6E8	GPU
Aurora	Transformer	0.25°	6 h	IFS-ANA	1.3E9	GPU
FourCastNet	Spectral Transf.	0.25°	6 h	ERA5	n/s	GPU, CPU
GraphCast	Graph-NN	0.25°	6 h	IFS-ANA	3.7E7	GPU, CPU
NeuralGCM	Hybrid	0.7°	1 h	ERA5	3.1E7	GPU, CPU
GenCast	Denoiser	0.25°	12 h	IFS-ANA	n/s	GPU

2. Experimental Design

In this section we describe the experimental setup for this study. In selecting the AI models, we chose examples of significant recent developments, each with a unique characteristic that stands out from the others. An overview over the main characteristics of each model is presented in Table 1. After introducing the models, we describe the construction of the initial conditions and the ensemble experiments.

2.1. Models

2.1.1. ICON

The ICON model (ICOsahedral Non-hydrostatic model; Zängl et al., 2015) is a complex "conventional" numerical weather prediction model, developed by the German Weather Service and the Max-Planck Institute for Meteorology. It applies a solver of the atmospheric partial differential equations and parameterization schemes to calculate sub-grid processes and processes not part of the fluid equation (e.g., radiation, moist processes). As a non-hydrostatic model based on a variable icosahedral grid it can be adapted to a wide range of horizontal resolutions.

2.1.2. Pangu

The model "Pangu Weather" (Bi et al., 2023) was the first in a set of AI-based weather models, which was able to slightly outperform the leading operational weather model at the European Centre for Medium-Range Weather Forecasts (ECMWF). Pangu is based on a vision transformer architecture with 2.6E8 learnable parameters, adapted for the atmosphere (3D earth specific transformer). It has been trained on ERA5 reanalysis data (Hersbach et al., 2020), takes only one time level as initial condition and has been trained on a single time step. This time step however varies from 1, 3, 6 to 24 hr, so in fact four different models have been trained. However, because the accuracy of the models gets worse as the step gets smaller, the models with smaller time steps are only intended to fill in the gaps ("hierarchical temporal aggregation"). Here we use only the 6 and 24-hr models, since most of the other AI models use a time step of 6 hr. The Pangu model was previously analyzed in our pilot study (Selz & Craig, 2023) and some of the results presented here are identical, but are included for direct comparison with the other models. We also investigate additional diagnostics which were not part of the earlier study.

2.1.3. Aurora

The Aurora model (Bodnar et al., 2025), like Pangu, employs a vision transformer architecture, but with 1.3E9 trainable parameters it is a much bigger model. It is designed as a foundation model with massive pretraining on diverse earth system data, including a mixture of forecasts, analysis data, reanalysis data and climate simulations. After this pretraining, a number of fine-tuned versions are available for different purposes. Here we use the 0.25° model version that is fine-tuned on the analyses from the Integrated Forecasting System (IFS) at ECMWF. Aurora applies a 6 hr time step, requires two initial states and the fine-tuning is done across two time steps.

2.1.4. FourCastNet

FourCastNet (Kurth et al., 2023) applies a Fourier Neural Operator (FNO) architecture, which implements global convolution via Fourier Transforms. It is essentially a transformer operating in spectral space. In this paper we use version 2 of FourCastNet which uses spherical FNOs (SFNOs), a generalization of FNOs on the sphere (Bonev et al., 2023). The model requires a single initial condition state, applies a 6-hr time step and is trained on ERA5 data. The first training phase has been performed on just one time step, followed by a "fine-tuning" on two consecutive time steps.

2.1.5. GraphCast

Instead of a transformer, the GraphCast model (Lam et al., 2023) implements a graph neural network on an icosahedral multi mesh with 3.7E7 trainable parameters. It uses a 6-hr time step and two consecutive states as initial condition and the training is applied over a 3-day lead time period. Originally, it was trained on ERA5 data, however in this study we use the version that was fine-tuned on the IFS operational analysis.

2.1.6. NeuralGCM

The NeuralGCM model (Kochkov et al., 2024) is a hybrid between a physics-based approach and machine learning. It consists of a dynamical core which is a standard spectral fluid solver based on the primitive equations to compute the general circulation of the atmosphere under the influence of gravity and the Coriolis force. This dynamical core is supplemented with a neural network which computes sub-grid processes, such as convection, radiation, cloud formation and precipitation. These correspond to the processes that are parameterized in conventional models like ICON. It is argued that these semi-empiric parameterizations introduce a lot of uncertainty and error into numerical weather models and are therefore suitable to be replaced by AI. In contrast to previous approaches, these AI-based parameterizations are trained "online" in NeuralGCM, that is, together with the dynamical core, to ensure proper interactions with the resolved motions. NeuralGCM is trained over 5 days on ERA5 reanalysis data, but the loss function includes an additional "sharpness loss" term, where excess smoothing of the forecast is penalized. For this study, we use the 0.7° version with 3.1E7 learnable parameters. This is the highest resolution available but has still about three times wider grid spacing than the other AI models. A time step of 1 hr is used, which is required for numerical stability of the dynamical core.

2.1.7. GenCast

The GenCast model (Price et al., 2025) aims at probabilistic weather forecasting by producing samples from an underlying distribution. It is a generative diffusion model based on a transformer architecture for the denoiser. It was trained originally on ERA5 reanalysis data, uses a 12 hr time step and takes in two initial conditions. Here we apply the 0.25° version with additional fine-tuning on IFS operational data.

2.2. Initial Conditions

As in Selz and Craig (2023), all experiments in this paper will be initialized on 26 June 2021, 00 UT. The case was selected because of strong convective activity over the North American continent at the initial time and during the following days. A very expensive convection-permitting ICON experiment has been conducted for this case, which will serve as a reference. The simulations and most of the diagnostics in this paper are global, hence the simulations will also contain maritime and wintertime conditions and the results show averages over many different weather regimes.

All experiments consist of 5-member ensemble simulations, started from the same 5-member ensemble of initial conditions. The initial conditions are based on the operational analysis from ECMWF with perturbations added to it to create the ensemble. The perturbations are retrieved from the ensemble of data assimilations (EDA) system at ECMWF (Isaksen et al., 2010), which applies perturbations to the observations and to the first-guess model to generate 50 samples of the initial condition uncertainty, from which we take the first 5 for this study. To simulate the butterfly effect, we drastically reduce the amplitude of the initial perturbations by a factor of 1/1,000 or equally 0.1% (other rescale factors will be applied in Section 3.6). After interpolation to the specific grid of each model, this set of 5 samples is used to initialize the model. Note that unlike the ECMWF operational ensemble (IFS-ENS), we do not add singular vectors to the initial conditions. For AI models that require two input states 6

or 12 hr apart, the earlier input is interpolated from the IFS operational analysis without any perturbation. This represents the ideal setup that everything is deterministic and perfectly known except for a tiny, "butterfly-like" perturbation at the initial time.

For the high-resolution reference simulation (ICON-2.5, see below) we slightly deviated from this approach and added the perturbations not to the IFS analysis but to a 24-hr forecast from the high-resolution ICON model, started 24 hr earlier. This ensures proper spin-up of small-scale motions that are not present in the IFS analysis.

Most global AI weather prediction models have been trained on the ERA5 reanalysis data set, although for some we apply versions which are fine-tuned on the IFS operational analyses (see Table 1). However, regardless of the data set used for training, in this study we start all simulations from the same initial conditions based on the IFS analysis plus rescaled perturbations. Slight performance penalties for models trained on ERA5 only might occur, but these are irrelevant for this study which considers perturbation growth rather than skill against observations.

2.3. Experiments

We apply the conventional, PDE-based ICON model to conduct reference experiments against which we will compare the AI experiments. To do so, we run ICON at four different resolutions, that is, 2.5 km (ICON-2.5), 20 km (ICON-20), 40 km (ICON-40) and 80 km (ICON-80). The ICON-2.5 experiment is identical to Selz and Craig (2023). It consists of convection-permitting simulations (i.e., the deep convection scheme has been turned off) and represents the best estimate of the butterfly effect currently feasible, since it explicitly resolves small-scale motions related to moist convection and cloud processes. Hence, this experiment will serve as the main reference. The ICON-20 experiment is included as another reference because it possesses a similar horizontal grid spacing to most of the AI models. However, some AI models have a lower grid resolution and most AI models have a much lower effective resolution, that is, they are overly smooth compared to their grid size. To address this fact we include two more experiments (ICON-40 and ICON-80) to create reference simulations with similar effective resolution (smoothness) for comparison. All ICON experiments other than ICON-2.5 employ a parameterization scheme for deep convection (Bechtold et al., 2008). The time step is adjusted according to the resolution and the vertical grid is kept unchanged. The ICON simulations have been computed on the ECMWF high performance computer ATOS. Output was retrieved on the native icosahedral grids of ICON and offline interpolated to a 0.25° latitude-longitude grid.

The AI model experiments have been conducted on NVIDIA H100 GPUs (Pangu, Aurora, GraphCast, FourCastNet, NeuralGCM, GenCast). However, we noticed that some AI models produce significantly different results when they were run on a CPU instead. We therefore conducted three additional AI experiments (FourCastNet-CPU, GraphCast-CPU and NeuralGCM-CPU), where x86-64-CPU processors were used to compute the simulations. We were not able to run Aurora and GenCast on CPUs and for Pangu it made no difference.

GenCast is a stochastic model and requires the specification of a seed value to specify a pseudo-random noise field from which a sample forecast is created. By providing the same initial state (the analysis) but specifying different seed values, an ensemble of forecasts can be generated to estimate the forecast uncertainty in an operational context. But because these uncertainties are much larger than the tiny, "butterfly like" perturbations we are interested in here, such a setup cannot be used for our purposes. Note that Kim et al. (2026) investigate operational uncertainty growth and not the "butterfly effect" as we understand it in this paper. Since we want to explicitly specify the initial condition uncertainty, we need to run GenCast in "quasi-deterministic" mode, which means keeping the seed value unchanged for each ensemble member.

There is also a stochastic version of NeuralGCM available at 1.4° resolution. Again, using a different seed value for each member would just reproduce the operational ensemble spread. Running the stochastic NeuralGCM in a "quasi-deterministic" mode (i.e., keeping the seed value unchanged) is possible, but didn't lead to any significant differences with respect to the deterministic version of NeuralGCM described above. Hence we did not include any runs from the stochastic NeuralGCM model in this paper.

All experiments are run up to a lead time of 3 days, which is sufficient to identify the 3 stages of uncertainty growth. Output frequency is 6 hr, except for GenCast, which uses a 12-hr time step. Additional simulations with a variety of different initial condition rescale factors and a much longer 45-day lead time have been conducted to

address the issue of convergence of the predictability time toward an intrinsic limit and will be described and evaluated in Section 3.6.

The diagnostics presented in this paper are all based on the horizontal wind (u, v) at 300 hPa. We use difference kinetic energy (DKE) as the main metric to quantify uncertainty, which has been used in previous studies on intrinsic predictability (e.g., Judt, 2018; Lorenz, 1969b; Selz et al., 2022) and which is defined as the ensemble variance of the horizontal wind, that is,

$$\mathrm{DKE} = \mathrm{var}(u) + \mathrm{var}(v). \tag{1}$$

For every experiment (AI and reference), the output is spatially interpolated to a 0.25° regular grid, if the output grid was different. Subsequently, the output is interpolated to a N360 Gaussian grid to compute the spherical harmonics transform. Interpolations and spherical harmonics transforms were computed with the Climate Data Operators (Schulzweida, 2024). If not stated otherwise, diagnostics are calculated globally.

2.4. Six Key Characteristics of the Butterfly Effect

We now lay out six key characteristics of the butterfly effect (i.e., uncertainty growth from tiny-amplitude initial perturbations) that are a consequence of the 3-stage error growth model and have been demonstrated in previous studies.

1. Fast initial growth rate. Large-amplitude uncertainties as found in current operational forecasts grow exponentially with a doubling time of about 1 day, but tiny initial perturbations will grow much faster, by several orders of magnitude in 12 hr, depending on their amplitude (Judt, 2018; Selz & Craig, 2015).
2. Adherence to precipitation. The uncertainties that quickly evolve from the tiny initial perturbations are spatially collocated with regions of precipitation and latent heat release, especially within areas of convection (Hohenegger & Schär, 2007; Selz & Craig, 2015).
3. Deviation from operational uncertainty pattern. Large-amplitude uncertainty growth in operational forecasts is related to the fastest growing modes of balanced motions on synoptic scales (singular vectors) and not directly associated with precipitation. Hence the spatial uncertainty pattern evolving from tiny initial perturbations differs from the uncertainty pattern in operational ensembles.
4. Initial growth on small scales. The fast initial growth of tiny perturbations is predominantly occurring on the convective scale or in its representation by the convection scheme on the model grid. Hence the uncertainty early in the simulation should peak at the convective scale or rather the smallest scales the model is able to adequately resolve (Selz & Craig, 2015; Selz et al., 2022).
5. Uncertainty growth from divergent wind. In the first day of the simulation, the uncertainties that developed on the convective scale spread out to larger scales by gravity wave propagation and through the outflow at the cloud tops. In contrast to balanced synoptic-scale dynamics these processes project strongly onto the divergent wind, hence the early uncertainty growth predominantly happens in the divergent component of the flow (Baumgart et al., 2019; Selz et al., 2022).
6. Existence of an intrinsic limit. While the growth rate of large-amplitude perturbations is mostly independent of the perturbation amplitude, once the initial perturbations get sufficiently small the growth rate will accelerate. In particular, the smaller the initial perturbations get, the faster the growth rate gets, leading to diminishing returns with respect to the forecast horizon and to the existence of an intrinsic predictability limit (Judt, 2018; Selz et al., 2022).

These six features will now be used to evaluate the simulation experiments from the AI models and the reference model. In each section, an objective metric will be defined to assess the manifestation of each of the listed features.

3. Results

3.1. Initial Growth Rate

The uncertainty time series of the different experiments measured by globally averaged DKE is presented in Figure 1. The DKE growth rates (inverse e-folding times) over the first 12 hr are shown in Figure 2(1).

10.1029/2025JH001180

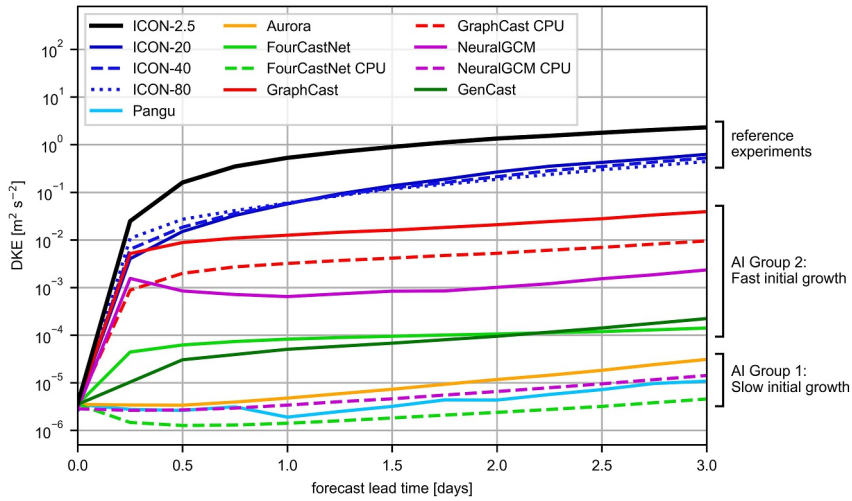

Figure 1. Globally integrated DKE time series over 3 days.

For tiny-amplitude initial perturbations, we expect the initial uncertainty growth rate to be very large (Section 2.4, 1). Indeed, the convection-permitting reference simulation (ICON-2.5) produces an uncertainty growth rate, which results in a DKE increase by about 5 orders of magnitude within the first 12 hr (Stage 1 of the error growth concept). The lower-resolution ICON simulations qualitatively resemble this behavior, with somewhat lower

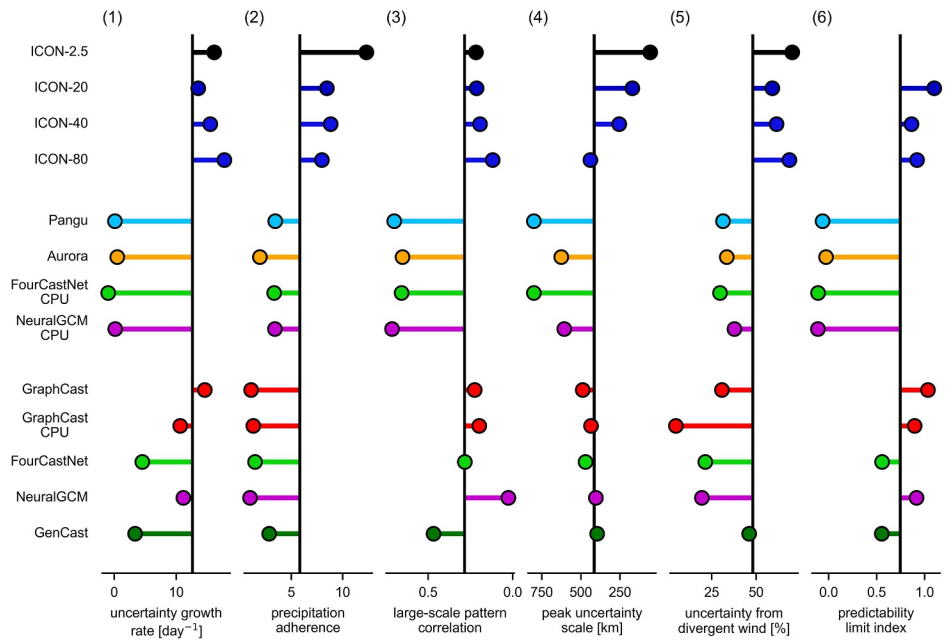

Figure 2. A set of metrics to represent six key characteristics of the butterfly effect. See text for further details on their computation. (1) The uncertainty growth rate measured by DKE over the first 12 hr. (2) The ratio of precipitation in areas of large uncertainty over precipitation in areas of small uncertainty at 12 hr. (3) The pattern correlation coefficient of the DKE field at 12 hr to the DKE field of the IFS-ENS. (4) The spatial scale at which the maximum uncertainty occurs at 12 hr. (5) The relative magnitude of the divergent wind contribution to uncertainty growth, estimated from a simplified PV diagnostic. (6) An index indicating the existence of a fundamental predictability limit when the initial condition uncertainty goes to zero. The experiments are sorted into three groups: (from top to bottom) Conventional reference experiments (ICON), AI experiments without an increased initial growth rate, AI experiments with an increased growth rate. The vertical line is introduced for visual aid and is computed as the average of the average of all reference experiments and the average of all AI experiments for each metric. Note that the *x*-axis orientation is reversed in (3, 4), so the reference experiment metrics always point to the right.

initial growth rates. The large initial growth rates decrease quickly and after about 48 hr the experiments transition to exponential uncertainty growth with a characteristic doubling time of about a day (Stage 3 of the error growth concept).

The AI models show a variety of different outcomes, which can be basically separated into two groups. The first group of experiments (Pangu, Aurora, FourCastNet-CPU, NeuralGCM-CPU) is completely unable to pick up the fast initial growth rate and right from the start produces the slower, synoptic growth rate or even a slight decrease of uncertainty, which is due to an initial loss of effective resolution. This is the behavior that was observed for Pangu in our initial study (Selz & Craig, 2023). However, not all AI models match this picture. A second group (FourCastNet, GraphCast, GraphCast-CPU, GenCast, NeuralGCM) shows at least some indication of an elevated initial growth rate, although with a wide range of values. The GraphCast experiment comes closest to the ICON reference simulations, followed by GraphCast-CPU and NeuralGCM. The experiments with FourCastNet and GenCast show only a rather weak initial acceleration of the uncertainty growth.

Interestingly, for most of the experiments that do produce an increased initial growth rate, the results are very sensitive to changing the hardware architecture. The FourCastNet and NeuralGCM experiments completely lose their fast initial growth when run on a CPU. GraphCast on CPU shows uncertainty diminished by a factor of about 5 compared to its GPU experiment. Since GenCast only ran on GPUs, we couldn't include a CPU experiment. These sensitivities to the hardware already suggest that the initial uncertainty increase in these models may be related to unphysical numerical noise.

From about 48 hr lead time on, all experiments (AI and reference) produce a similar growth rate, that is, the lines in Figure 1 are approximately parallel. This indicates that all AI models, regardless of their ability to generate a fast initial growth rate, are able to represent the slower synoptic-scale growth process which dominates in operational weather prediction.

3.2. Adherence to Precipitation

After addressing the globally integrated uncertainty growth, we now inspect the spatial uncertainty patterns that evolved from the different experiments. Figure 3 shows the normalized DKE structures after 12 hr over Europe and Asia, together with the DKE from the operational IFS ensemble (IFS-ENS) and accumulated precipitation, derived from the deterministic IFS forecast.

Although some differences are present, all four ICON simulations show the same characteristic DKE pattern, with large amplitudes found over East Asia, the Western Pacific and Indian Ocean. The individual peaks differ significantly between the ICON-2.5 simulation, with explicit convection, and the lower resolution experiments ICON-20, ICON-40, and ICON-80, with parameterized convection, but the envelopes of the structures are clearly similar. We expect that early uncertainty growth is spatially collocated with precipitation (Section 2.4, 2). For ICON, this is confirmed by the figure, where the precipitation field is shown in the last row.

In contrast to ICON, this close relationship to precipitation is not reproduced by any of the AI models. They either largely reproduce the growth patterns of the operational IFS ensemble (Pangu, Aurora, FourCastNet-CPU, NeuralGCM-CPU; see below) or they produce a large amount of evenly distributed and seemingly unphysical noise (FourCastNet, GraphCast, GraphCast-CPU, NeuralGCM, GenCast). To demonstrate these visual impressions in a more quantitative way, we averaged the 12-hr accumulated precipitation (again estimated by the IFS deterministic forecast) over the areas where the experiment's DKE field at 12-hr lead time exceeded the 75%-quantile and over the area where the DKE field was below the 25%-quantile, respectively. After that, we divided these two numbers; the higher the index is, the more the DKE patterns are related to precipitation. The values are given in Figure 2(2) and confirm the visual impression from the figure.

3.3. Correlation to Operational Uncertainty Pattern

We revisit the maps in Figure 3, this time comparing them with the uncertainty patterns arising from large-amplitude initial perturbations, as represented by the operational IFS ensemble (IFS-ENS). We expect the uncertainty structures evolving from tiny and large-amplitude perturbations to be different (Section 2.4, 3). Indeed for the ICON reference simulations, the uncertainty structures are mostly different from those of the IFS-ENS. This is confirmed by the spatial correlation coefficient between the experiments and the IFS-ENS, which is given in Figure 2(3), showing correlations of the order of only 0.2.

10.1029/2025JH001180

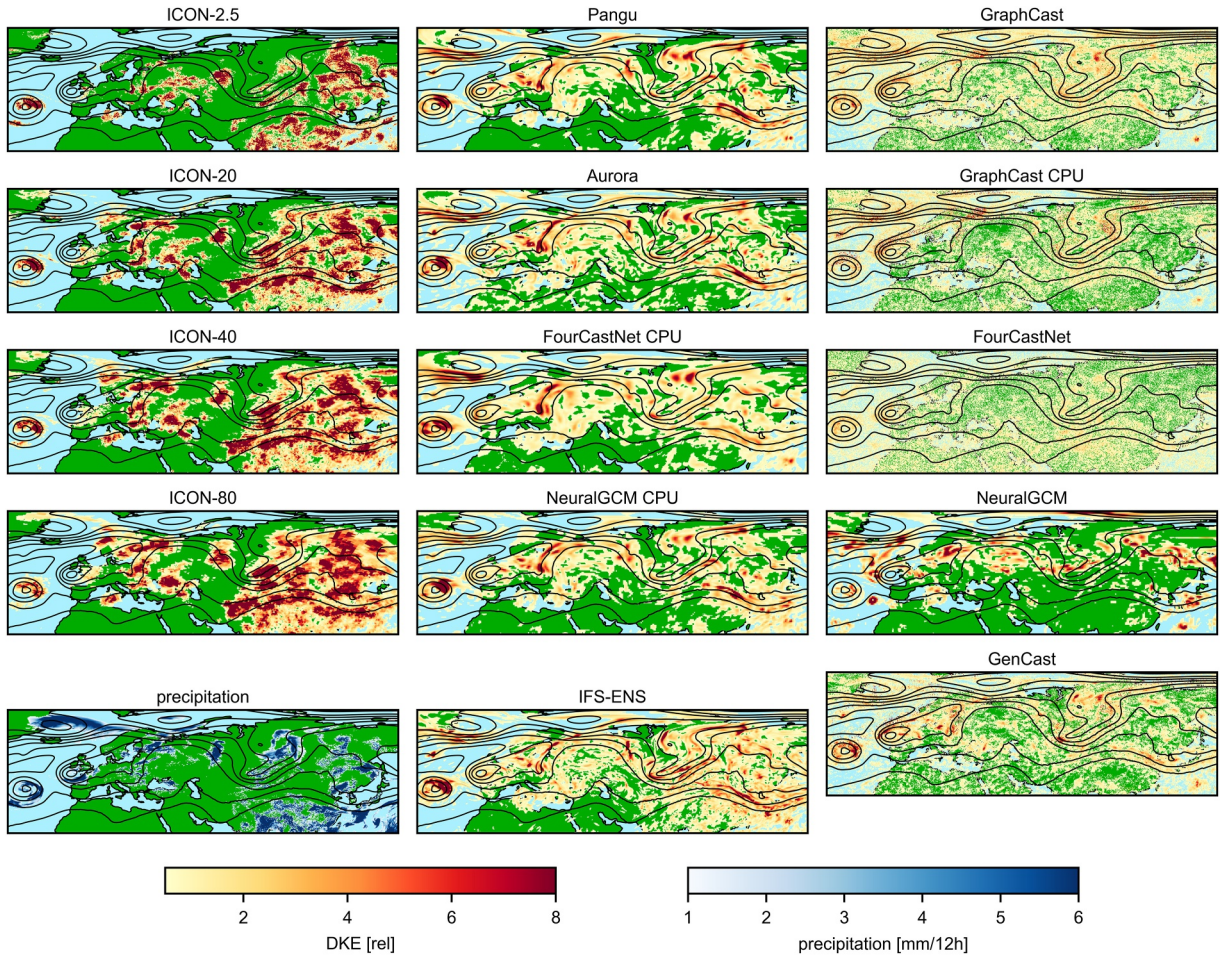

Figure 3. Spatial DKE patterns at 12 hr lead time for different experiments over Europe and Asia. The black lines show the ERA5 300 hPa geopotential as reference. The DKE amplitudes are normalized to globally average to one. In addition, the DKE pattern of the IFS operational ensemble and the 12-hr accumulated precipitation from the IFS deterministic forecast are shown. Green and light blue background colors show the land-sea mask for reference.

For the AI experiments, the picture is more diverse. The first group of AI experiments (Pangu, Aurora, FourCastNet-CPU, NeuralGCM-CPU), which did not show fast initial growth, produces uncertainty patterns which are very similar to each other and to the pattern of the IFS-ENS. Consequently, the correlation coefficient plotted in Figure 2(3) is high (>0.5). The most striking difference is more small-scale substructures in IFS-ENS, due to its higher (effective) resolution. The second group of AI experiments (FourCastNet, GraphCast, GraphCast-CPU, NeuralGCM, GenCast) generated noise and the correlations to the IFS-ENS pattern are highly variable, depending on the distribution of this noise.

3.4. Spatial Scale of the Uncertainty Peak

Before analyzing the spatial scale of the evolving uncertainties, it is important to characterize the spectral properties of the models themselves, that is, their effective resolution. As already mentioned in the introduction, the grid size of a model (the numerical resolution) does not necessarily correspond to the size of the smallest simulated features (the effective resolution), since model output can be smooth over many adjacent grid points. Especially AI-based models have been reported to be overly smooth, in part because of their choice of the loss function (Selz et al., 2025). A standard way to assess the effective resolution of a simulation is to compute its kinetic energy (KE) spectrum, which is shown in Figure 4 for each experiment. To highlight differences between the experiments, all KE spectra are plotted relative to the KE spectrum of the IFS Control simulation (IFS-CTL), which is the unperturbed member of the ensemble.

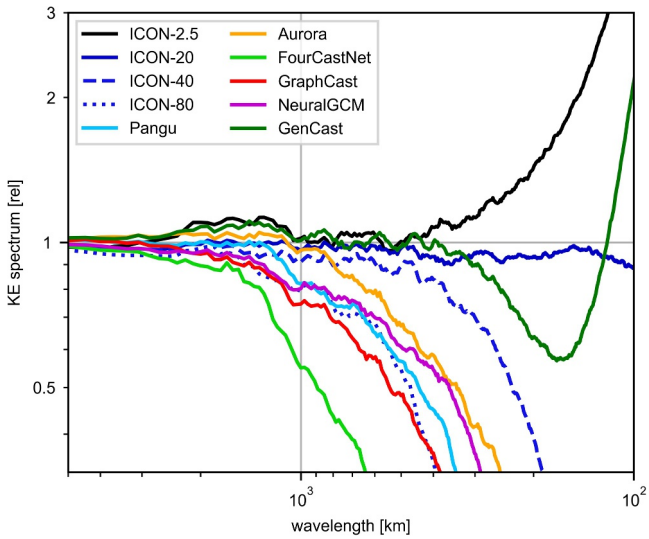

Figure 4. KE spectra of different experiments, averaged over lead times between 48 and 72 hr, normalized with the IFS-CTL spectrum.

For the ICON reference experiments, the kinetic energy on small scales is clearly related to the grid size. The ICON-2.5 experiment shows higher levels of kinetic energy than the IFS-CTL on scales smaller than 400 km, where the observed energy spectrum transitions from a steep k^{-3} slope to a shallower $k^{-5/3}$ dependence (Nastrom & Gage, 1985). ICON-2.5, having a roughly 8 times finer grid than the IFS-CTL is better able to simulate these small scales. The ICON-20 experiment produces a KE spectrum similar to the IFS-CTL, while ICON-40 and especially ICON-80 produce much lower kinetic energy, with some reduction even on the synoptic scale.

Among the AI experiments, FourCastNet is the smoothest, already underestimating the kinetic energy on the 2,000-km scale with a very sharp decline below 1,000 km. The other AI experiments start to significantly differ from the IFS-CTL KE spectrum on scales between 1,000 and 500 km, with the only exception being GenCast, which is able to maintain its kinetic energy down to the 300-km scale. Note that we have not included the CPU-based experiments in Figure 4 since the differences in kinetic energy relative to the GPU-based simulations are tiny. The impact of numerical noise produced by GPUs only shows clearly when differences are taken (see below).

The ICON-20 simulation has a grid size which is comparable to most of the AI experiments (the ones that use a 0.25° grid), but its kinetic energy on small scales exceeds the kinetic energy of the AI models by far. This low effective resolution of the AI models is due to a combination of low model capacity, coarser internal representation and the smoothing effect of mean-squared-error-based training (Selz et al., 2025). For this reason we added the lower-resolution ICON experiments to include a more similar reference. The figure reveals that the ICON-40 experiments is still much sharper, but the ICON-80 experiment produces a kinetic energy spectrum that broadly compares to the KE spectrum of most of the AI experiments.

Having discussed these background spectra, we now focus on the spectra of the developing uncertainties in terms of difference kinetic energy (DKE), which are given in Figure 5. The butterfly effect initially triggers fast uncertainty growth mostly related to convection, implying that peak uncertainty should shift from the 500-km scale in the initial condition perturbations to the convective scale, or rather the smallest scales that the model can resolve at the 12-hr lead time (Section 2.4, 4). The scale of the uncertainty peak at 12 hr is also shown in Figure 2(4). In addition, as already discussed earlier, there should initially also be a very large increase in amplitude due to the fast initial growth rate. After that, the uncertainty should continue to grow more slowly and move upscale. We can clearly see this happening in all the reference ICON simulations, indicated by the connected dots in Figure 5, which mark the maxima of the DKE spectra. We can also clearly see the effect of lowering the resolution in ICON, namely the initial jump of the uncertainty to small scales gets less pronounced at lower resolution because these scales cannot be (accurately) resolved.

In contrast to ICON, most of the AI experiments do not reproduce this behavior and if they do, this again is linked to unphysical noise: In the first AI-experiment group (Pangu, Aurora, NeuralGCM-CPU and FourCastNet-CPU), there is no shift of the peak uncertainty to smaller scales; the peak instantly shifts to larger scales. The spectra also confirm that there is no significant amplitude increase, only at 72 hr there is a small increase in uncertainty amplitude related to the large-scale growth effects these models reproduce. The second group of AI experiments (FourCastNet, NeuralGCM, GenCast, GraphCast, Graphcast-CPU) does show some shift of the peak uncertainty to smaller scales and also some increase in amplitude. But this again is a consequence of the generation of numerical noise. In the DKE spectra, the noise can be identified by its stagnating amplitude growth between 12 and 72 hr on small scales. This stagnation indicates saturation of the uncertainty, however, a physically meaningful uncertainty should saturate at twice the level of the background spectrum, which is at least a factor of 10 above this "fake" saturation level. It can also be seen from the spectra, that the projection of this noise onto the larger synoptic scales does grow further between 12 and 72 hr and the peak uncertainty shifts to about 1,000 km at 72-hr lead time.

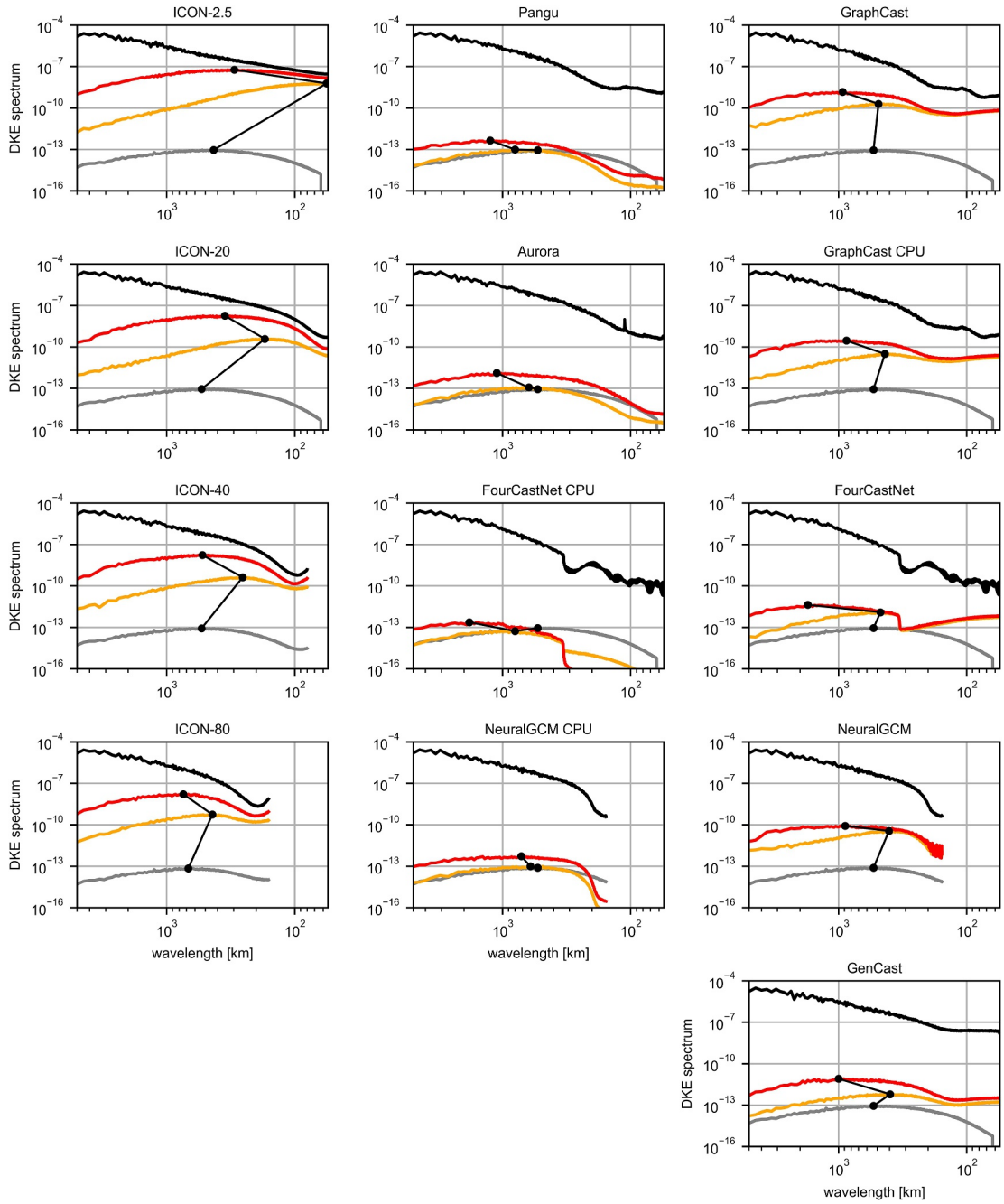

Figure 5. Spectra of DKE for different experiments. The colored lines show the DKE spectrum at different lead times: 0 hr (gray), 12 hr (yellow), 72 hr (red). The dots indicate the DKE maxima. The black line shows the background KE spectrum at 72 hr lead time. The spectra are only shown up to the Nyquist wavelength of the model grid.

3.5. Contribution of Divergent Wind

After having investigated amplitude, pattern and scale, we now attempt to relate the uncertainty growth to physical processes. For tiny-amplitude perturbations, the uncertainty growth is expected to occur predominantly in the divergent part of the flow (Section 2.4, 5). To estimate this, we consider the growth of ensemble variance of potential vorticity (PV), which can be decomposed into contributions from the divergent and rotational parts of the wind and also various diabatic processes, represented by tendencies from parameterization schemes in a PDE-based model (Baumgart et al., 2019; Baumgart & Riemer, 2019).

However, to be able to apply this diagnostic to the current study, significant simplifications have to be made. First, there is no equivalent to tendencies from parameterization schemes in most AI models. The exception is NeuralGCM, where combined tendencies from the learned parameterization could be derived in principle, but an interface to these tendencies has not been provided. Hence the diagnostic has to be limited to contributions from divergent and rotational wind, which can easily be computed for all experiments via a Helmholtz decomposition (Dawson, 2016). Second, the calculation of PV is difficult for AI models, since it requires knowledge of the stratification, that is, the vertical temperature gradient, which due to the coarse vertical spacing can only be calculated with significant error. In addition, most AI models are prone to much larger errors when approaching the tropopause and the stratosphere. For these reasons, we decided to assume a constant, uniform stratification, which essentially reduces the PV-based diagnostic to a vorticity-based diagnostic on the 300 hPa isobaric level.

Other than that, we follow Baumgart and Riemer (2019) and Selz et al. (2022) and compute the tendency of the ensemble enstrophy variance due to the divergent and the rotational wind, respectively, and at 12-hr and 24-hr lead time. The tendencies are spatially integrated over the midlatitudes (defined as 40° to 60° north and south), averaged over the two time steps and the ratio of the divergent tendency to the total tendency (the sum of divergent and rotational tendency) is taken. The resulting value then resembles a "divergent wind index", indicating which fraction of the uncertainty was generated by the divergent wind in the first 24 hr. Note also that the diagnostic is calculated on a 2.5° grid to allow accurate gradient calculations, including the experiments with a grid spacing larger than 0.25°.

The index is plotted in Figure 2(5). As expected, all ICON reference experiments show a much larger contribution to uncertainty growth from the divergent wind than from the rotational wind (index >50%). In contrast, the AI experiments clearly show the opposite situation, that is, the rotational wind dominates uncertainty growth (index <50%). The group of experiments that did not produce unphysical noise all show an index value between 30% and 40% (Pangu, Aurora, FourCastNet-CPU, NeuralGCM-CPU), while the index value varies substantially for the noise-producing experiments (FourCastNet, NeuralGCM, GenCast, GraphCast, GraphCast-CPU), likely due to different spatial characteristics of the noise (compare Figures 3 and 5).

3.6. Indication of a Fundamental Predictability Limit

An important consequence of the butterfly effect is limited intrinsic predictability, which means that no matter how small the initial condition uncertainty gets, the forecast horizon cannot be extended beyond a certain lead time (Section 2.4, 6). This is a direct consequence of the increasing initial growth rate when the initial uncertainty decreases, leading to diminishing returns with respect to the predictability time. Although it can be inferred whether a certain model indicates the existence of a predictability limit from the initial growth rates shown in Figure 1, we will investigate this property explicitly by varying the initial condition uncertainty over three orders of magnitude and deriving the predictability time from the data.

To do so, we conducted additional simulations by again computing five member ensembles for each AI and reference experiment, but starting them from 10 different initial condition perturbation amplitudes (ranging from 100% to 0.1% relative to the EDA perturbations) and running them for 45 days. To reduce the data volume, the output of these runs is interpolated to a regular 2.5° grid. Note however, that we were not able to include a convection-permitting ICON-2.5 experiment here because of the immense computational cost. Hence for this set of experiments, ICON-20 has the highest spatial resolution and serves as the main reference.

The predictability time for each ensemble simulation is defined as the time it takes until the ensemble spread represented by globally integrated DKE at 300 hPa reaches a certain threshold. The selection of this threshold is arbitrary and reflects the level of uncertainty a particular user is willing to tolerate before the forecast is considered useless. Here, the DKE threshold is defined as 25% of the DKE saturation level, and we estimate this saturation level by averaging the DKE values of the ICON-20 reference experiment started from 100% initial condition uncertainty between day 30 and day 45.

Figure 6 shows the resulting predictability time and the initial growth rate as a function of initial condition uncertainty for each experiment. In addition, a quantitative metric to identify a model's ability to indicate the existence of an intrinsic limit can be defined by comparing the slopes on both sides of the predictability time curve (Figure 6, bottom). Let m_0 be the slope derived from a linear regression of the leftmost 4 points (i.e., 100%–10% initial condition uncertainty) and m_1 be the slope analogously derived from the rightmost 4 points (i.e., 1%–0.1%).

10.1029/2025JH001180

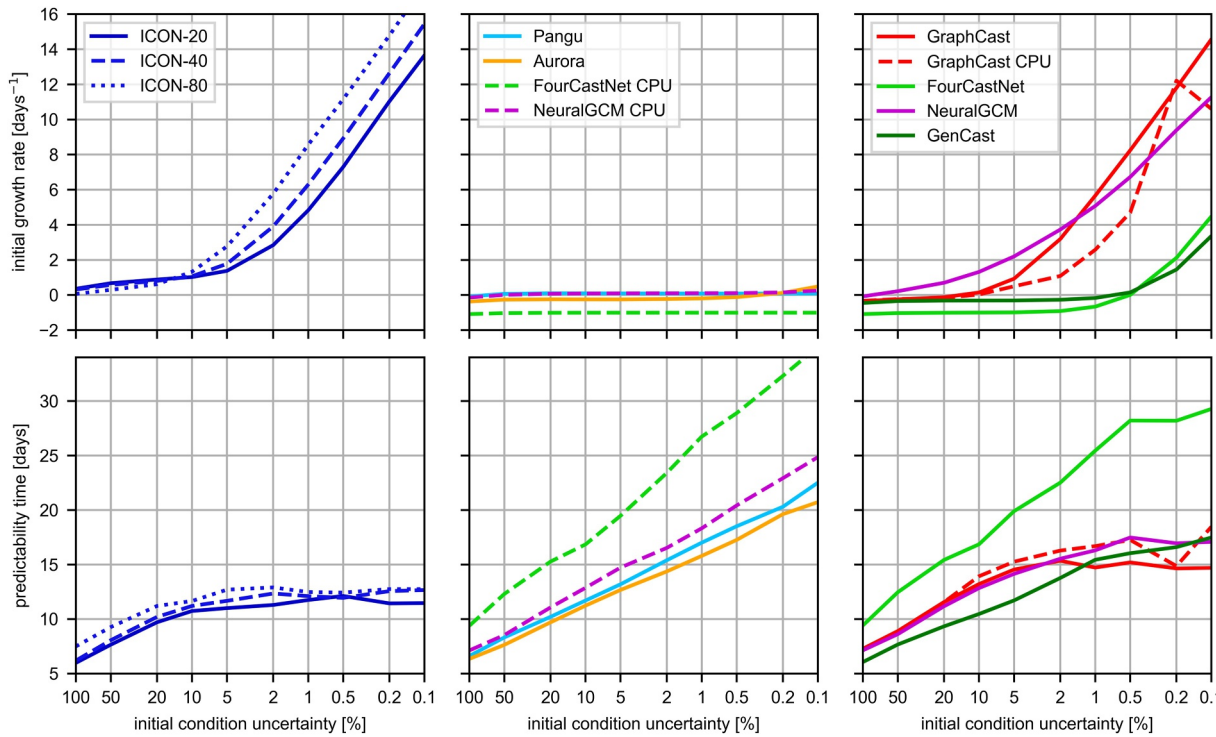

Figure 6. (Upper row) Initial DKE growth rate (first 12 hr) and (lower row) predictability time (i.e., time for DKE to reach a threshold) over the amplitude of the initial condition perturbation.

With this we define an intrinsic "predictability limit index" as $1 - m_1/m_0$, which is shown in Figure 2(6). The index is 1 if the slope on the right side of the curve is 0, indicating a leveling off to a horizontal plateau and hence the presence of a predictability limit. The index is 0 if the slope on the right and left side of the curve is the same, which indicates unlimited predictability.

The ICON reference experiments show an approximately constant initial growth rate from the 100%-initial uncertainty level until about 10%, resulting in a linear increase in predictability time (Figure 6 left). Note that the x-axis is log-scaled. From the 10% level on down to 0.1% this is followed by a continuous increase of the initial growth rate and consequently diminishing additional improvement of the predictability time. Despite some quantitative differences between the different resolutions of the ICON experiments, they all consistently indicate an intrinsic limit of predictability, which is about 12–13 days for the chosen metric and threshold. This is also largely consistent with previous results (Selz et al., 2022).

While the ICON reference experiments show a constant (i.e., amplitude-independent) initial growth rate and thus a linear increase in predictability time only for large initial uncertainties, the AI-model experiments of the first group (Figure 6 center) show a constant growth rate and linear predictability time increase for all the initial perturbations, including 0.1%. Hence their intrinsic limit index is close to zero (Figure 2(6)) and these models indicate an unlimited predictability of the atmosphere, that is, forecasts can be extended indefinitely by "just" making the initial condition uncertainty smaller.

The AI experiments of the second group (Figure 6 right) do however show deviations from a constant initial growth rate, with the predictability time leveling off for small initial uncertainties. Hence, these experiments to varying degrees do indicate a limited atmospheric predictability, their intrinsic limit index is close to one. For FourCastNet the leveling off only happens from the 0.2% level on and the significance of this result is unclear. The same is true for GenCast, which shows a slightly increasing growth rate and reduced predictability time from the 0.5% level on. NeuralGCM and GraphCast-CPU more clearly show a predictability limit, which is about 17 days and the GPU version of GraphCast indicates a limit of about 15 days. Compared to the ICON reference experiments, the estimated predictability limit is too long and the uncertainty level at which the limit is reached is too low.

Consistent with our previous findings, the set of experiments that show at least some indication of an intrinsic limit is identical to the set of experiments that produced unphysical noise. The diminishing returns that occur at some point in these experiments indicate that the noise which is mostly generated on small scales is able to affect larger synoptic-scale and planetary-scale motions eventually. However, the weak contributions of the divergent wind show that the propagation of the uncertainty to larger scales is not achieved via physical processes of divergent outflow and geostrophic adjustment like in the reference experiments, but by synoptic-scale growth of the part of the noise spectrum that projects onto those scales initially (compare Figure 5).

4. Summary and Discussion

As a follow-up from our pilot study on the Pangu model (Selz & Craig, 2023), we revisited the issue of uncertainty growth from tiny initial condition perturbations (the "butterfly effect"), testing a variety of recently published AI weather models with different sizes, training and architectures, including models based on vision transformers, spectral transformers, graph neural networks, hybrid and generative approaches. In addition, we considered more reference simulations from a conventional physics-based model using a variety of resolutions, from cutting-edge convection-permitting resolution, through simulations with a similar grid size, to simulations with an effective resolution (smoothness) similar to the AI models.

We evaluated the models by investigating six key characteristics of the butterfly effect, identified in previous studies and reproduced by the high-resolution ICON-2.5 experiment. This convection-permitting experiment is considered the most accurate representation of reality available and therefore serves as the main reference. The six key features are (a) a fast initial growth rate, (b) initial uncertainty growth adhering to the spatial locations of precipitation but (c) different from operational uncertainty patterns, (d) dominant initial growth on small scales and (e) driven by the divergent wind, which leads to (6) an intrinsic limit. While even the lower-resolution ICON experiments were able to at least qualitatively reproduce all of these butterfly features, none of the AI models could reproduce all of them in a consistent way.

More specifically, the AI experiments could be separated into two groups with respect to their reaction to the tiny perturbations. The experiments of the first group (Pangu, Aurora, FourCastNet-CPU, NeuralGCM-CPU) showed no evidence of a butterfly effect and generated uncertainty growth very similar to that of large-amplitude uncertainties in operational ensembles. They produced opposite results compared to the ICON reference simulations on every tested measure: The initial growth rate was low, the growing uncertainty patterns were not related to precipitation but to operational uncertainties and there was no initial shift of the peak uncertainty to smaller scales. In addition, the uncertainty growth was mostly driven by the rotational wind and there was no indication that predictability is limited.

The second group of experiments (GraphCast, GraphCast-CPU, FourCastNet, NeuralGCM, GenCast) did qualitatively agree with the reference experiments on some of the tested measures. Most noteworthy, these experiments did show an increased initial growth rate and indicated limited predictability to some degree. Still, no experiment agreed with the reference experiments on all of the butterfly features to paint a physically consistent picture. The adherence of uncertainty growth to precipitation was absent and uncertainty growth was still mostly driven by the rotational wind. The generation of unrealistic numerical noise could be identified as the main reason for the accelerated initial uncertainty growth which affects the synoptic scales mostly by projection, in turn triggering large-scale growth and limiting the estimated predictability from these experiments. For FourCastNet and NeuralGCM this noise could be completely eliminated by conducting these experiments on a CPU instead of the intended GPU; for GraphCast the noise level was significantly reduced on a CPU. The noise is likely caused by numerical errors specific to GPUs: The reduced numerical precision which the AI models apply on GPUs leads to truncation errors that may act as random perturbations larger than the tiny initial perturbations. In addition, GPUs can introduce non-deterministic numerical variability, which can cause run-to-run noise even with identical initial conditions. These errors are usually neither noticeable nor relevant in operational settings that involve much larger uncertainties.

The AI models we have been testing vary in their size, training data sets, architecture and effective spatial resolution. However, apart from the generation of noise, none of these different properties appeared to improve the representation of the butterfly effect. The massive pretraining of the very large foundation model Aurora, even on forecast data showed no effect. The spectral transformer FourCastNet, which could perhaps represent scale interactions more accurately also did not lead to a butterfly effect representation. The hybrid model NeuralGCM

which uses AI only for the parameterizations could have more accurately represented latent heat release in convection as the driving process of fast initial uncertainty growth, but no effect on the butterfly metrics was seen. Finally, the generative AI model GenCast produces a much more realistic kinetic energy spectrum and can better represent small-scale features, but this too did not help to improve the butterfly effect representation.

Despite their differences in architecture, size and training, a common property of all the AI models considered here is that they are trained or fine-tuned on analysis data (operational IFS analyses or ERA5 reanalyses). The analysis states in the training data set differ much more from the unknown true initial state than the 0.1% perturbations we tested in this paper. In fact, they differ by 100%, assuming that the initial condition estimation in the ECMWF EDA is correct. Such large-amplitude errors grow directly on synoptic scales and in the balanced, rotational component of the atmospheric circulation. In addition, except for GenCast, all AI experiments presented here use a similar loss function, based on root-mean-square or mean-absolute errors. In the limit of infinite model capacity and perfect training, the loss function and the training data will essentially determine the model's behavior. Resolution, design and architecture are of minor importance. The ideal model will infer the 100% initial uncertainty from the training data to provide an ensemble-mean prediction from that distribution, at least over the short lead-time interval that is included in its loss function (Selz et al., 2025). It follows that a perfectly trained deterministic AI model *should* produce similar forecasts when given slightly perturbed initial conditions. This critical importance of the training data also explains why even the hybrid NeuralGCM model was not able to reproduce the fast initial uncertainty growth. The online-trained "parameterizations" of NeuralGCM act as corrections to the dynamical core to optimize the same target (loss function) as the other AI models. Accordingly, the observed behavior of the first group of AI models (the "non-noise" models) is perfectly consistent with the information provided by the training data and the loss function.

From the point of view of simulating the butterfly effect as a physical phenomenon however, the outcome of our study is mostly negative. This raises the question what the implications are for operational weather forecasting, where these models are now widely used. Although not tested specifically in this paper, most AI models appear to be able to reproduce the fastest growing modes on the large scales (singular vectors), which will lead to generally realistic uncertainty growth in the context of operational midrange weather prediction (e.g., Baño-Medina et al., 2025). In this context there is currently no indication that the inability to reproduce the butterfly effect has any negative consequences: As illustrated by the results shown here (see also Selz et al., 2022), tiny "butterfly-like" perturbations grow differently compared to the growth of the large uncertainties that operational forecasts are dealing with. Also the unrealistic numerical noise that some of the AI models produced becomes irrelevant in the presence of these much larger uncertainties. However, it is possible that upscale uncertainty propagation is much more important during specific and rare atmospheric flow patterns, which are often discussed in the context of forecast dropouts. Such patterns include heavy continental convection (e.g., over North America) and extra-tropical transition of tropical cyclones, which can cause significant impacts downstream and reduce predictability (Lillo & Parsons, 2017; Rodwell et al., 2013).

Weather prediction models are used not only for operational forecasting. They are also used as research tools to address fundamental scientific questions about the atmosphere. By exploring intrinsic predictability and uncertainty growth, our study exemplifies that conventional and AI-based models can yield fundamentally different answers despite generating forecast products of similar skill. We demonstrated numerous physical inconsistencies in the AI-model results and large discrepancies among them, leading us to conclude that, for this topic, the conventional models are more reliable (see Vonich & Hakim, 2025 for an alternative view). More broadly, the issue of trust in scientific findings derived from AI models will become increasingly important: AI systems remain largely black boxes, and their low cost and differentiability may enable future studies whose results cannot directly be verified against conventional approaches or observations.

Setting these more general concerns aside, the question remains of how one could enable AI models to correctly and physically consistently reproduce the butterfly effect. While there are still many open questions on the details, the basic requirement is that the butterfly effect, that is, the fast growth of tiny uncertainties on the convective scale with subsequent upscale propagation, must be contained in some way in the training data. As explained above, this is not the case when analyses data sets are used, since they possess much larger errors compared to reality. One obvious way to teach stochastic AI models the butterfly effect is to train them on ensemble data generated from tiny initial perturbations like the ICON reference experiments in this study using, for example, the continuous ranked probability score (CRPS). Such a model would then be started from a single initial state and

different random seed numbers would be used for different members to generate an ensemble, which then should approximate the key properties of the training ensemble. However, such a model would only work specifically for the specified initial perturbation amplitude in the training data and would likely have very limited use for practical weather forecasting.

What it would take for deterministic models to exhibit the butterfly effect is more difficult to answer. The purely data-driven AI models analyzed in this paper can be regarded as a revival of the idea of forecasting by finding close analogs from the past. Lorenz (1969a) estimated that it requires an impractically long time for analogs to occur, depending on the desired proximity of the analogue. While Lorenz was concerned with hemispheric analogs, the AI models are likely able to look for local analogs and generalize them in a clever way, making these models successful in weather prediction based on training data sets of a few decades only. However, to be able to learn the butterfly effect from data, much closer local analogs need to be present in the training data set and the model needs to have sufficient capacity to pick them up. A high-resolution, convection-permitting reanalysis data set is likely still insufficient since even the high-resolution analysis states will suffer from large-amplitude errors relative to the underlying truth, although it might help produce more realistic short-range forecasts. Training from a single, long, high-resolution simulation (a "nature run") from a conventional model which can be used to represent an exactly known alternative truth might theoretically work. However, it seems unlikely that this approach would be practically feasible, due to requirements for simulation length and model capacity. And in any case, the AI would be learning to emulate the conventional model, rather than the real atmosphere.

Conflict of Interest

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

Availability Statement

The models used in this study are publicly available and can be retrieved from the following URLs. ICON (https://www.icon-model.org), Pangu (https://github.com/198808xc/Pangu-Weather), Aurora (https://microsoft.github.io/aurora/intro.html), FourCastNet (https://catalog.ngc.nvidia.com/orgs/nvidia/teams/modulus/models/modulus_fcnv2_sm?version=v0.2), NeuralGCM (https://github.com/neuralgcm), GraphCast and GenCast (https://github.com/google-deepmind/graphcast). ERA5 reanalysis data is available via the Copernicus Climate Data Store (Hersbach et al., 2017). ECMWF analyses, EDA perturbations and forecasts are available via the MARS-archive at ECMWF (https://apps.ecmwf.int/mars-catalogue/?class=od, restricted access). The simulation data from the AI models and ICON that has been evaluated in this paper can be retrieved from Selz and Craig (2026).

Acknowledgments
The work of Tobias Selz was supported by the German Research Foundation (DFG) under project number 548044620 ("What is the relevance of the butterfly effect for practical weather prediction?"). The use of ECMWF's computing and archive facilities is gratefully acknowledged. We are grateful to the anonymous reviewers for their careful reading of the manuscript and for their constructive comments that substantially improved the paper. Open Access funding enabled and organized by Projekt DEAL.

References

Baño-Medina, J., Sengupta, A., Doyle, J. D., Reynolds, C. A., Watson-Parris, D., & Monache, L. D. (2025). Are AI weather models learning atmospheric physics? A sensitivity analysis of cyclone Xynthia. *npj Climate and Atmospheric Science*, *8*(1), 92. https://doi.org/10.1038/s41612-025-00949-6

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

Baumgart, M., Ghinassi, P., Wirth, V., Selz, T., Craig, G. C., & Riemer, M. (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

Baumgart, M., & Riemer, M. (2019). Processes governing the amplification of ensemble spread in a medium-range forecast with large forecast uncertainty. *Quarterly Journal of the Royal Meteorological Society*, *145*(724), 3252–3270. https://doi.org/10.1002/qj.3617

Baumgart, M., Riemer, M., Wirth, V., Teubler, F., & Lang, S. (2018). Potential vorticity dynamics of forecast errors: A quantitative case study. *Monthly Weather Review*, *146*(5), 1405–1425. https://doi.org/10.1175/mwr-d-17-0196.1

Bechtold, P., Köhler, M., Jung, T., Doblas-Reyes, F., Leutbecher, M., Rodwell, M. J., et al. (2008). Advances in simulating atmospheric variability with the ECMWF model: From synoptic to decadal time-scales. *Quarterly Journal of the Royal Meteorological Society*, *134*(634), 1337–1351. https://doi.org/10.1002/qj.289

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

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Allen, A., & Brandstetter, J. (2025). A foundation model for the Earth system. *Nature*, *1–8*.

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. In *International Conference on machine learning* (pp. 2806–2823).

Dawson, A. (2016). Windspharm: A high-level library for global wind field computations using spherical harmonics. *Journal of Open Research Software*, *4*(1), 31. https://doi.org/10.5334/jors.129

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) Data Store (CDS)*. https://doi.org/10.24381/cds.143582cf

Hohenegger, C., & Schär, C. (2007). Predictability and error growth dynamics in cloud-resolving models. *Journal of the Atmospheric Sciences*, *64*(12), 4467–4478. https://doi.org/10.1175/2007jas2143.1

Isaksen, L., Bonavita, M., Buizza, R., Fisher, M., Haseler, J., Leutbecher, M., & Raynaud, L. (2010). Ensemble of data assimilations at ECMWF.

Judt, F. (2018). Insights into atmospheric predictability through global convection-permitting model simulations. *Journal of the Atmospheric Sciences*, *75*(5), 1477–1497. https://doi.org/10.1175/jas-d-17-0343.1

Kim, H., Ryu, J., Son, S.-W., Jeong, J.-H., Kim, H., & Yoon, J.-H. (2026). A spectral test of the butterfly effect and physical consistency in the diffusion-based Gencast's ensembles. *npj Climate and Atmospheric Science*.

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

Kurth, T., Subramanian, S., Harrington, P., Pathak, J., Mardani, M., Hall, D., et al. (2023). Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive Fourier neural operators. In *Proceedings of the platform for Advanced Scientific Computing Conference* (pp. 1–11).

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

Lillo, S. P., & Parsons, D. B. (2017). Investigating the dynamics of error growth in ECMWF medium-range forecast busts. *Quarterly Journal of the Royal Meteorological Society*, *143*(704), 1211–1226. https://doi.org/10.1002/qj.2938

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, *20*(2), 130–141. https://doi.org/10.1175/1520-0469(1963)020⟨0130:DNF⟩2.0.CO;2

Lorenz, E. N. (1969). Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric Sciences*, *26*(4), 636–646. https://doi.org/10.1175/1520-0469(1969)26<636:aparbn>2.0.co;2

Lorenz, E. N. (1969). The predictability of a flow which possesses many scales of motion. *Tellus*, *21*(3), 289–307. https://doi.org/10.3402/tellusa.v21i3.10086

Nastrom, G., & Gage, K. (1985). A climatology of atmospheric wavenumber spectra of wind and temperature observed by commercial aircraft. *Journal of the Atmospheric Sciences*, *42*(9), 950–960. https://doi.org/10.1175/1520-0469(1985)042<0950:acoaws>2.0.co;2

Palmer, T., Döring, A., & Seregin, G. (2014). The real butterfly effect. *Nonlinearity*, *27*(9), R123–R141. https://doi.org/10.1088/0951-7715/27/9/r123

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., et al. (2025). Probabilistic weather forecasting with machine learning. *Nature*, *637*(8044), 84–90. https://doi.org/10.1038/s41586-024-08252-9

Rodwell, M. J., Magnusson, L., Bauer, P., Bechtold, P., Bonavita, M., Cardinali, C., et al. (2013). Characteristics of occasional poor medium-range weather forecasts for Europe. *Bulletin of the American Meteorological Society*, *94*(9), 1393–1405. https://doi.org/10.1175/bams-d-12-00099.1

Schulzweida, U. (2024). CDO user guide. *Zenodo*. https://doi.org/10.5281/zenodo.7112925

Selz, T. (2019). Estimating the intrinsic limit of predictability using a stochastic convection scheme. *Journal of the Atmospheric Sciences*, *76*(3), 757–765. https://doi.org/10.1175/jas-d-17-0373.1

Selz, T., Bruinsma, W., Craig, G. C., Markou, S., Turner, R., & Vaughan, A. (2025). *On the effective resolution of AI weather prediction models*. Authorea Preprints. https://doi.org/10.22541/essoar.174139239.94807670/v1

Selz, T., & Craig, G. C. (2015). Upscale error growth in a high-resolution simulation of a summertime weather event over Europe. *Monthly Weather Review*, *143*(3), 813–827. https://doi.org/10.1175/mwr-d-14-00140.1

Selz, T., & Craig, G. C. (2023). Can artificial intelligence-based weather prediction models simulate the butterfly effect? *Geophysical Research Letters*, *50*(20), e2023GL105747. https://doi.org/10.1029/2023gl105747

Selz, T., & Craig, G. C. (2026). Data for "can AI-based weather prediction models simulate the butterfly effect? The role of architecture and implementation." [Dataset]. *LMU Munich*. https://doi.org/10.57970/1csc0-cwr90

Selz, T., Riemer, M., & Craig, G. C. (2022). The transition from practical to intrinsic predictability of midlatitude weather. *Journal of the Atmospheric Sciences*, *79*(8), 2013–2030. https://doi.org/10.1175/jas-d-21-0271.1

Vonich, P. T., & Hakim, G. J. (2025). Testing the limit of atmospheric predictability with a machine learning weather model. *arXiv preprint arXiv: 2504.20238*. https://doi.org/10.48550/arXiv.2504.20238

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

Zhang, F., Bei, N., Rotunno, R., Snyder, C., & Epifanio, C. C. (2007). Mesoscale predictability of moist baroclinic waves: Convection-permitting experiments and multistage error growth dynamics. *Journal of the Atmospheric Sciences*, *64*(10), 3579–3594. https://doi.org/10.1175/jas4028.1

Zhang, F., Sun, Y. Q., Magnusson, L., Buizza, R., Lin, S.-J., Chen, J.-H., & Emanuel, K. (2019). What is the predictability limit of midlatitude weather? *Journal of the Atmospheric Sciences*, *76*(4), 1077–1091. https://doi.org/10.1175/jas-d-18-0269.1