Geophysical Research Letters®

RESEARCH LETTER

10.1029/2023GL105747

Key Points:

- Current artificial-intelligence-based models cannot simulate the butterfly effect and incorrectly suggest unlimited atmospheric predictability
- Their error growth rate and structure remain similar to synoptic-scale error growth regardless of the amplitude of the initial perturbation
- Synoptic-scale error growth from current levels of initial condition uncertainty appears mostly realistic, except for a short initial decay

Correspondence to:

T. Selz,
tobias.selz@lmu.de

Citation:

Selz, T., & Craig, G. C. (2023). Can artificial intelligence-based weather prediction models simulate the butterfly effect? *Geophysical Research Letters*, *50*, e2023GL105747. https://doi.org/10.1029/2023GL105747

Received 31 JUL 2023
Accepted 22 SEP 2023
Corrected 4 NOV 2023

This article was corrected on 4 NOV 2023. See the end of the full text for details.

Can Artificial Intelligence-Based Weather Prediction Models Simulate the Butterfly Effect?

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

[1]Deutsches Zentrum für Luft- und Raumfahrt, Oberpfaffenhofen, Germany, [2]Ludwig-Maximilians-Universität, München, Germany

Abstract We investigate error growth from small-amplitude initial condition perturbations, simulated with a recent artificial intelligence-based weather prediction model. From past simulations with standard physically-based numerical models as well as from theoretical considerations it is expected that such small-amplitude initial condition perturbations would grow very fast initially. This fast growth then sets a fixed and fundamental limit to the predictability of weather, a phenomenon known as the butterfly effect. We find however, that the AI-based model completely fails to reproduce the rapid initial growth rates and hence would incorrectly suggest an unlimited predictability of the atmosphere. In contrast, if the initial perturbations are large and comparable to current uncertainties in the estimation of the initial state, the AI-based model basically agrees with physically-based simulations, although some deficits are still present.

Plain Language Summary Even if perfect observations and models were available, the time interval for which weather forecasts can be accurate is limited. This limit is related to fundamental physical characteristics of the earth's atmosphere, which make small errors grow very fast and spread out, a feature known as the butterfly effect. In this article, we test if an artificial intelligence-based weather prediction model is able to reproduce this butterfly effect. Therefore, we computed several weather forecasts that differed only very slightly in their starting conditions. We find, that in contrast to standard weather forecasting models, the initial difference grow only slowly in the AI-based model and there is no indication of a butterfly effect at all. This provides an example of how machine learning models can fail to reproduce a fundamental physical principle, even though they can accurately mimic many observed behaviors.

1. Introduction

The "butterfly effect" refers to a well-known and unfortunate property of the atmospheric circulation: Tiny uncertainties or errors in the initial conditions are rapidly amplified, creating a fundamental, intrinsic predictability limit for weather forecasting that cannot be overcome. This limit was first identified by Lorenz (1969), and has since been studied in detail using various methods, including complex numerical weather prediction models with very high resolutions or stochastic parameterizations (e.g., Judt, 2018; Selz, 2019; Zhang et al., 2019). The fundamental reason for the existence of this limit is scale interactions (Lorenz, 1969; Palmer et al., 2014), especially originating from the convective scale, where highly nonlinear dynamics, enforced by latent heat release can lead to very rapid error growth (Selz & Craig, 2015b; Zhang et al., 2007).

If the amplitude of the initial perturbations is sufficiently small, its spatial structure (i.e., the scale of the "butterfly") is no longer important and tiny errors on any scale will lead to rapid growth, saturation of small-scale errors and subsequent upscale error propagation (Durran & Gingrich, 2014; Sun & Zhang, 2016). However, larger-amplitude uncertainties on synoptic and planetary scales grow by a different mechanism, sometimes called up-amplitude growth, where errors grow exponentially in time at the same rate for all scales until saturation (Durran & Gingrich, 2014; Rotunno & Snyder, 2008). On average, the initial uncertainty that is present in current operational weather prediction systems is large enough that this latter mechanism dominates (Selz et al., 2022). However, the transition to the former process would occur already if the initial uncertainty was reduced to 10%–20% of its current level. This raises the possibility that for some weather situations, the butterfly effect may already significantly limit forecast skill, which is a topic of active research (e.g., Craig et al., 2021).

In current practice, weather forecasts are computed based on a set of partial differential equations (PDEs), which are mathematical formulations of the laws of physics. Those PDEs are then discretized, approximated and optimized using various numerical methods and usually solved on massive-parallel computer architectures. Recently, a novel approach is been pursued, where weather forecasts are computed with artificial-intelligence (AI) based, data-driven methods. Those methods apply deep neural networks, that have been trained on a series of historical atmospheric states (e.g., Bi et al., 2023; Lam et al., 2022; Pathak et al., 2022; Weyn et al., 2019), usually obtained from reanalysis data sets like ERA5 (Hersbach et al., 2020). Neural networks estimate the future atmospheric state by interpolating and combining developments that happened in the past without any direct knowledge of physical laws or constraints. Recent results have shown forecast skill comparable, or even superior to conventional forecast models (e.g., Bi et al., 2023; Lam et al., 2022), with the AI models having the huge advantage that, once trained, they require much less computational effort to compute a weather forecast. This could help to reduce cost and energy consumption in weather forecasting and/or free resources to extend ensemble sizes or data assimilation.

In this paper, we investigate the ability of a state-of-the-art AI-based model (Pangu) to simulate the butterfly effect, that is, very fast error growth from very small-amplitude initial condition perturbations. We compare these results to simulations with a state-of-the-art numerical prediction model based on PDE discretizations (ICON), including a simulation with convection-permitting resolution. We also investigate whether the AI model is able to accurately simulate error growth from current estimates of the initial condition uncertainty.

2. Experiments

2.1. The AI-Based Model Pangu

As a representative of the class of data-driven AI-based models, we apply “Pangu-Weather” (Bi et al., 2023), which has very recently been published and is free to use for research purposes. It has been shown to produce slightly better deterministic forecasts than the leading operational weather prediction model (IFS), evaluated with standard metrics like root-mean-square error or anomaly correlation with respect to the ERA5 reanalysis. Pangu consists of a 3D deep neural network that has been trained with 39 years of ERA5 data. The model state of Pangu consists of 13 pressure levels (from 1,000 to 50 hPa) with 5 upper-air variables (horizontal wind, temperature, geopotential and specific moisture), which are complemented by 4 surface variables (10-m horizontal wind, 2-m temperature and mean sea-level pressure). Additional variables like precipitation are not computed. All variables are defined on a regular 0.25°-lat-lon grid. These variables are propagated forward in time, where 4 different networks are provided for 4 different time steps (1 hr, 3 hr, 6 hr, 24 hr). Longer time steps produce better forecasts, hence a “hierarchical temporal aggregation” technique is used, where first the longest time step (24 hr) is applied consecutively, followed by the next shorter time step and so on, until the desired time resolution is reached. For example, to produce a set of hourly forecasts out to three days, the 24-hr network is used to produce forecasts at 24 hr, 48 and 72 hr, the 6-hr time step network is then used to fill in times 6 hr, 12 hr, 18 hr, 30 hr, 36 hr, etc., followed by application of the 3-hr and then the 1-hr time step networks so that a forecast is available for every hour.

2.2. The PDE-Based Model ICON

The Pangu simulations will be compared to simulations with ICON (ICOsahedral Non-hydrostatic model; Zängl et al., 2015), which is a complex PDE-based numerical weather prediction model. Hence it solves a discretized and approximated version of mathematical equations that describe the atmosphere. ICON consists of a non-hydrostatic dynamical core and operates on an icosahedral grid with terrain-following height levels. Processes that are not properly resolved with the grid resolution or that are not part of the fluid equations are parameterized, which means estimated with simplified and approximated methods. Such processes include convection, as well as cloud microphysics, radiative interactions, turbulence, gravity wave drag and interactions with the surface boundary.

2.3. Initial Conditions

For the experiments in this paper, we focus on simulations initialized at 26 June 2021, 00 UT. This case was originally selected because it showed significant amounts of continental summertime convection over North America during the following days, which could lead to rapid error growth from small uncertainties. However, since the simulations are global, the atmosphere at that time also contains maritime and wintertime conditions, and the global diagnostics applied here will be averages over many different weather systems. We use the operational analysis from the European Centre for Medium-Range Weather Forecasts (ECMWF) to initialize both Pangu and ICON. Initializing Pangu with the ERA5 reanalysis instead, which it was trained on, led only to minor and insignificant differences.

To provide an estimate of the initial condition uncertainty in current weather prediction systems, initial perturbations are retrieved from the ensemble data assimilations (EDA) system at ECMWF (Isaksen et al., 2010). The EDA system uses perturbed observations and a model uncertainty representation scheme to estimate the initial condition uncertainty, sampled with a 50 member ensemble. Those perturbations where interpolated to the Pangu and ICON grids, rescaled (see below) and added to the analysis state to create an ensemble of initial conditions.

2.4. Experiments

For this paper, we conducted five different experiments, which will be labeled as Pangu-100%, Pangu-0.1%, ICON-LR-100%, ICON-LR-0.1% and ICON-HR-0.1%. The first term describes the model that has been used (Pangu or ICON). While the spatial resolution of Pangu is fixed, with ICON we simulate two different resolutions: LR (low resolution) and HR (high resolution). For the low resolution runs, ICON is set up with about 20 km grid spacing (R2B7-grid), which is a horizontal resolution similar to that of Pangu. This resolution requires a time step in the dynamical core of 36 s, since ICON explicitly resolves horizontal-propagating sound waves. The high resolution experiments use a 2.5 km grid spacing (R2B10-grid) and a time step of 4.5 s, which allows a convection-permitting simulation with the parameterization scheme for deep convection turned off. This latter experiment is believed to provide the best estimate of convective-scale error growth and the butterfly effect currently available. It explicitly resolves the convective motions and does not any more rely on a convection parameterization, which has been shown to slow down error growth (Selz & Craig, 2015a).

The percentage factor (100% or 0.1%) indicates a rescaling of the initial condition perturbations derived from the EDA system. 100% means we took them as they are without any changes and they represent an estimation of the current level of initial condition uncertainty. 0.1% means we reduced their amplitude by a factor of 1,000, which leads to a very small uncertainty in the initial condition ensemble. These experiments will represent "butterfly"-like perturbations and provide estimates of the intrinsic limit. They are also sometimes called "identical twin" experiments. Note, that the initial perturbations do not include singular vectors and the models used in this study are deterministic, and do not contain any stochastic parameterization or representation of model uncertainty. These experiments are therefore suited to estimate basic atmospheric error-growth properties in a perfect-model context and not designed to produce reliable probabilistic forecasts.

While running Pangu is very cheap and running ICON at the low resolution is pretty affordable, the high resolution ICON simulations at global convection-permitting resolution are very expensive. As a result, we only could simulate 5 ensemble members out of 50 and with three days of integration time. We also abstained from producing a high-resolution ICON simulation for the 100% initial perturbations, since previous studies suggest that the small-scale processes that the HR simulation would represent more accurately are not crucially relevant there (Selz et al., 2022). All experiments were computed on CPUs, the ICON experiments on the Atos computer at ECMWF. The computational costs were 16 core hr for each Pangu experiment, 2,900 core hr for each ICON-LR experiment and 1,300,000 core hr for the ICON-HR experiment.

The setup of the HR simulations is further complicated by the fact that the analyses from ECMWF (which provide the initial conditions) do not contain convective-scale motions, because their resolution is similar to that of Pangu or ICON-LR. Therefore, the ICON-HR experiment needs to spin up those small scales first and starting this experiment just from the perturbed analysis would not produce the correct perturbation growth. To allow the small scales to spin up, we started one simulation of ICON-HR one day earlier (from the 25 June 2021, 00 UT analysis) and ran it for 24 hr. The complete model state was then written to disk, the rescaled EDA perturbations were added and the 5-member ensemble was run for three days lead time. We consider the slight difference in the initial state at the 26 June 2021, 00 UT much less significant than investigating error growth in a HR simulation where the small-scales are not present in the initial conditions.

2.5. Diagnostics

In this study we focus on difference kinetic energy (DKE) on 300 hPa as our main diagnostic. For an ensemble, DKE is defined as

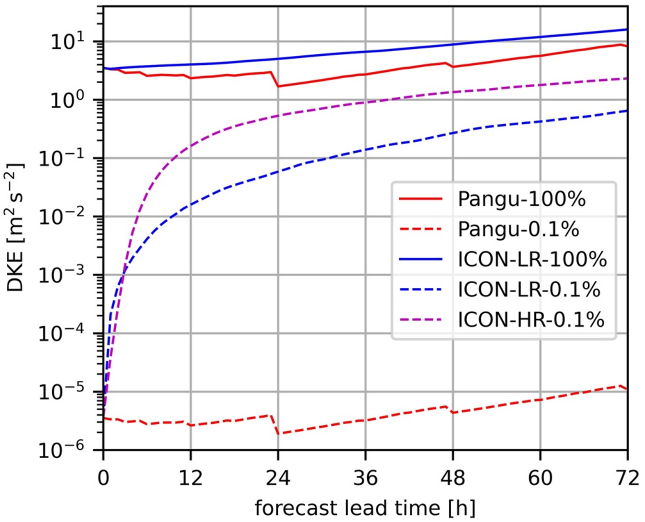

Figure 1. Globally-averaged difference kinetic energy (DKE) as defined by Equation 1 on 300 hPa over time for the different experiments (hourly output time step).

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

with the horizontal wind components u and v and the variance taken over the ensemble dimension. DKE has been frequently used to study error growth and intrinsic predictability. Note, that DKE is a metric to diagnose error growth in the sense of spread or variance growth in the ensemble and not errors with respect to observations or analyses. For all experiments, DKE is calculated based on wind data on a regular 0.25° grid and with an hourly output time step. While Pangu directly outputs this data, it is interpolated from the ICON grids using conservative remapping.

In addition, we calculated a spectral representation of DKE, since this is directly related to the kinetic energy (KE) spectrum of the atmosphere or the model and provides additional insights. To do so, we used the Climate Data Operators (Schulzweida, 2022) to compute spherical harmonics expansions of divergence and vorticity from the gridded horizontal wind, which first had to be interpolated to a Gaussian grid (N360). Further details of how to compute spectra of KE and DKE can be found in Augier and Lindborg (2013) and Selz et al. (2022).

3. Results

First, we looked at the errors in the ensemble mean of the simulations with respect to the ERA5 reanalysis. This is only been done to check that all models are implemented correctly, since deterministic forecast accuracy is not a focus of this paper. The root-mean-squared error (RMSE) of the ensemble-mean zonal wind at 300 hPa and 72 hr forecast lead time is similar for all experiments and lies between 4.6 and 4.9 m s^{-1}, with Pangu being marginally better. This is comparable to the average error that is shown in Bi et al. (2023) for the zonal wind on 500 hPa and indicates that all models used here are able to produce similar forecast quality with respect to upper-level winds.

3.1. Time Series of DKE

To study error growth, we start with investigating the 72 hr time series of globally averaged DKE for the different experiments, which is shown in Figure 1.

Note that, although DKE is globally averaged, this average largely favors the mid-latitudes because of the concentration of kinetic energy there. The y-axis is log-scaled, so exponential growth can be identified with a straight line. Consistent with earlier studies (Selz et al., 2022), we see exponential growth right from the start for the 100% ICON experiment with a growth rate of about 1.7 day^{-1}, which is characteristic of the synoptic-scale dynamics of the atmosphere. The 100% Pangu experiment shows a similar growth rate, starting from 24-hr lead time (2.2 day^{-1}). Two anomalies are apparent in the Pangu curve: In the first day, there is a decrease in the ensemble spread, which in the end leads to a 1.9 times lower DKE at 72 hr compared to ICON. Also, Pangu shows discontinuities when switching to a model with a different time step. The 24-hr network in particular shows the largest initial decrease in DKE compared to the shorter time-step networks.

At the intrinsic limit (0.1% experiments), the two ICON experiments show the expected very large initial growth rates of $\mathcal{O}\left(10^{20}\ \mathrm{day}^{-1}\right)$ during the first 3 hr. This leads to a very fast saturation of the errors at small scales (see Section 3.3), which quickly slows down the error growth, until the characteristic synoptic-scale growth rate is reached after about 48-hr lead time. A further reduction of the initial condition uncertainty would lead to even larger initial growth rates, leading to very little reduction in DKE at later times and keeping predictability limited. As anticipated, the lower resolution model appears to underestimate the error growth from small perturbations. This underestimation leads to about a factor 4 reduction in DKE at 72 hr. Nevertheless, both ICON experiments show the fast initial growth characteristic of the butterfly effect and clearly indicate the limited intrinsic predictability of the atmosphere.

In contrast, the 0.1% Pangu experiment essentially reproduces the error growth properties of the 100% experiment, with only a very slight increase of the growth rate. The two Pangu lines are very similar and just shifted

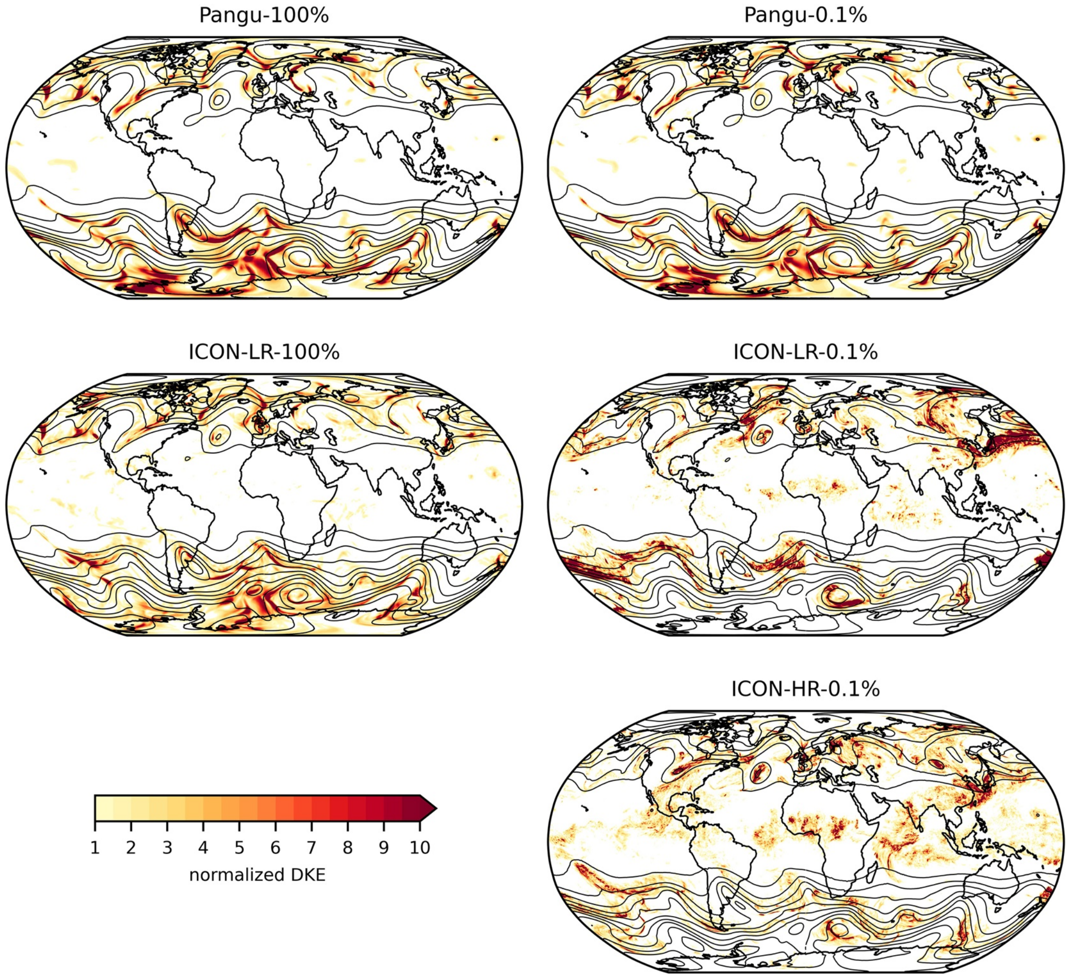

Figure 2. Global maps of normalized DKE on 300 hPa after 72-hr lead time. The thin black lines show the 300 hPa geopotential of the ensemble mean for reference (linespacing 1,500 $m^2 s^{-2}$).

vertically by 1000^2, the squared rescale factor of the initial perturbations (since DKE is a squared quantity). The increase in DKE after 72 hr compared to the initial conditions is still only a factor of 3.1 (0.1% experiment), similar to 2.3 for the 100% experiment. Over the three days, this leads to an underestimation of the DKE by five orders of magnitude compared to ICON. The almost constant error growth rates in the simulations with Pangu would incorrectly indicate an unlimited predictability of the atmosphere and no presence of a butterfly effect.

3.2. Spatial Structures of DKE

Next, we consider the spatial patterns of DKE that have evolved from the initial condition uncertainty after 72 hr (Figure 2). For the plot, the DKE is normalized so that the area-weighted mean equals one for a better comparison of the structures (the corresponding amplitudes have been shown in Figure 1).

Consistent with the similar growth rates, Pangu-100% and Pangu-0.1% also generate almost identical DKE structures, regardless of the amplitude of the initial condition uncertainty. The DKE maps of Figure 2 are visually difficult to differentiate and their spatial correlation coefficient equals 0.74. In contrast, there is much less structural agreement between the 100% and 0.1% experiments simulated with ICON: Although some common “hotspots” are visible, the different processes that drive the error growth in the two experiments (Selz et al., 2022) lead to largely disjunct structures after 72 hr. ICON-LR-100% is only correlated by 0.21 to ICON-LR-0.1% and by 0.14 to ICON-HR-0.1%.

For the 100% initial perturbations, ICON and Pangu show a reasonable level of agreement between the DKE structures after 72 hr (correlation 0.56). This is noteworthy, since recently Rodwell and Wernli (2023) pointed out that even among state-of-the-art operational models there is a lot of discrepancy in ensemble spread growth.

In contrast, for the butterfly-like perturbation, Pangu-0.1% does not show any agreement with the structures computed with ICON (correlation 0.06 to ICON-LR-0.1% and 0.05 to ICON-HR-0.1%), again indicating the inability of Pangu to simulate the butterfly effect.

3.3. Spectra of DKE

As final diagnostic we consider global spectra of 300 hPa kinetic energy and difference kinetic energy (Figure 3).

Similar to the globally-averaged DKE timeseries, they largely indicate mid-latitude conditions. Black lines show the background KE spectra of the simulations (ensemble mean of the member spectra), taken at 72 hr lead time. The initial condition perturbations have very little influence on them, since the simulations still display the same basic state and, in general, the background spectrum is largely a climatological feature of the atmosphere or rather the model. Hence, we mostly see differences between Pangu and ICON and also between the two ICON resolutions.

During the first 24 hr, Pangu looses a significant amount of energy compared to the initial spectrum on scales below about 500 km, where it is not able to maintain the expected −3 slope. This loss of energy on small scales is most pronounced with the first call of the 24-hr model, which significant smooths the initial state. However, after that, the spectrum remains almost stable, that is, successive calls of the Pangu 24-hr model do not lead to a further decay of kinetic energy. Hence we can assign an effective resolution of about 500 km to Pangu, which equals $20\Delta x$. For ICON-LR, the smoothing happens around the 200-km scale, which indicates an effective resolution of about $10\Delta x$. To determine the effective resolution of the ICON-HR simulations, a higher output resolution would have been required, but it likely also similar to $10\Delta x$.

The DKE spectrum of the initial condition perturbation is shown in gray and is basically identical for all the simulations, except of course for the rescale factor of 1000^2 and some very minor differences due to interpolation to the icosahedral grid of ICON and back. This initial DKE distribution is given by the EDA system of ECMWF and peaks at around 500 km. Scales below about 200 km are already saturated (see ICON-LR-100%), meaning there is no information about those scales from the data assimilation system. Starting from the initial uncertainty, we see a clear signature of synoptic-scale exponential error growth (equidistant lines) in the ICON-LR-100% experiment, with the energy maximum slowly moving to larger scales as more and more small scales saturating. The Pangu-100% experiment shows a similar behavior, except in the first 24 hr, where (as described above) background kinetic energy is removed from scales below 500 km. This also leads to a stagnation of error growth, even at large scales. After that and for scales larger than around 1,000 km, growth rates and spectral characteristics appear realistic and are very similar to ICON. In particular, Pangu does not consecutively smooth out scales that have lost predictability.

As seen previously in the spatial patterns, Pangu-0.1% shows similar behavior to Pangu-100%, with the difference kinetic energy reduced by the rescale factor 1000^2. This leads to a huge gap compared to the background spectrum after 72 hr and no indication of a butterfly effect. In contrast, ICON-LR-0.1% and even more so ICON-HR-0.1% show extremely large initial growth rates (as already discussed in Figure 1), but also an initial and almost instantaneous downscale propagation of the energy peak to smaller scales, another signature of the butterfly effect (Durran & Gingrich, 2014; Lorenz, 1969; Selz et al., 2022). This originates from decorrelation of small-scale structures by small differences in the large-scale advection and also by fast and highly non-linear error growth in regions of moist convection. After that initial downscale propagation, the errors grow slowly back upscale, finally transitioning to the synoptic-scale growth regime with similar growth rates as in the 100% experiments.

4. Discussion

The main advantage of the new AI-based models over standard PDE-based models is their very low computational cost, and it is frequently stated that this opens up the opportunity to create many more ensemble members (e.g., Bi et al., 2023). Indeed, large ensembles would greatly reduce sampling uncertainty and could provide much more reliable forecasts of extreme event probabilities (Tempest et al., 2023). Such big ensembles will however only make sense, if the error growth properties of the model are realistic. Here, the AI-based Pangu-Weather model that we tested is able to reproduce basic error growth properties when started from 100% initial condition perturbations, although some significant shortcomings were also observed. However, it fails completely when the initial condition uncertainty is small and is not able to simulate any indication of the butterfly effect or any indication of accelerated growth rates and intrinsically limited predictability.

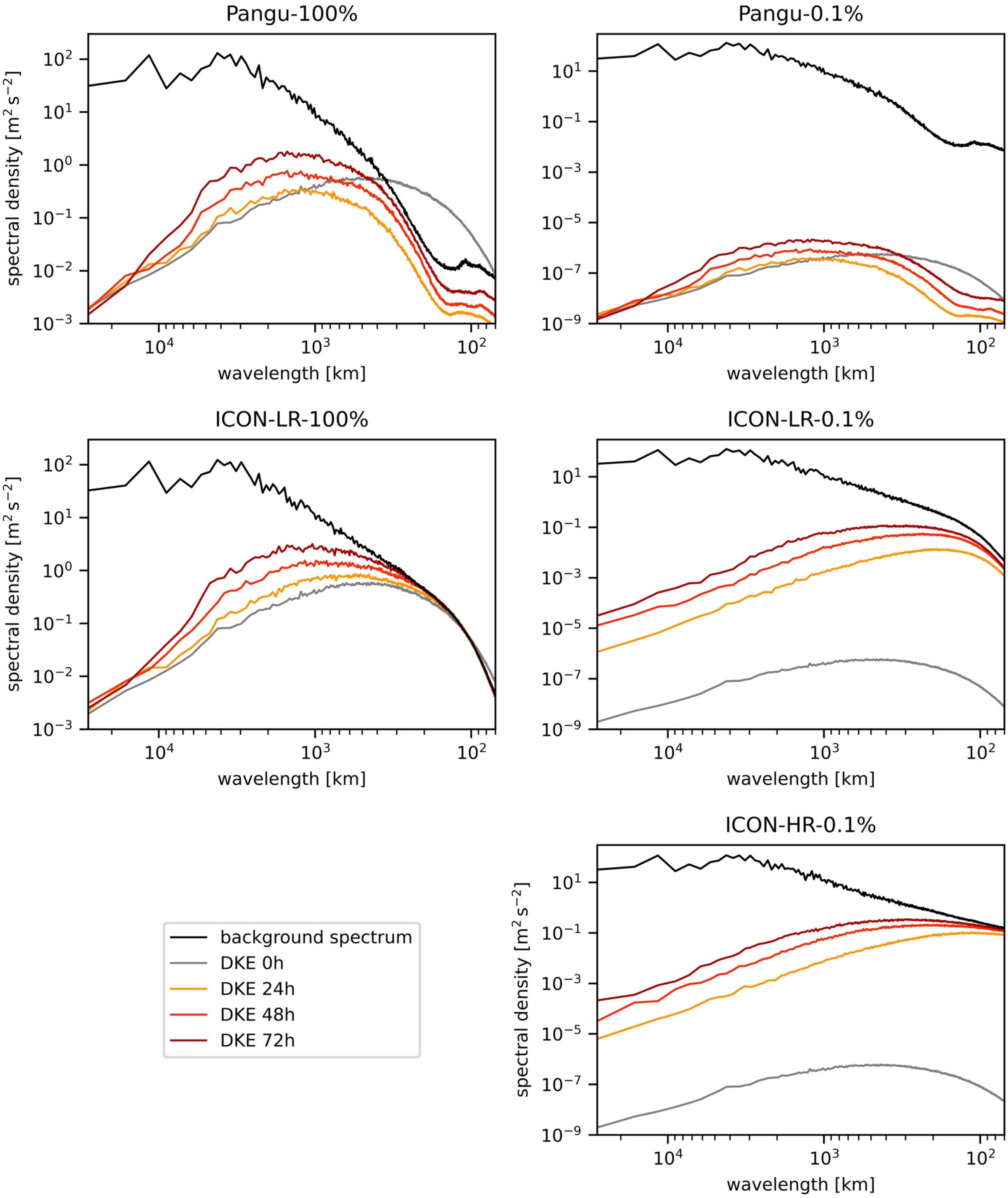

Figure 3. Spectra of 300 hPa kinetic energy (KE, black line, evaluated at 72 hr lead time) and different kinetic energy (DKE, colored lines). The DKE spectra are multiplied by 0.5 so that they match the background spectrum at saturation.

This failure to simulate the butterfly effect is perhaps not surprising for AI-based models: Although the intrinsic predictability limit is a basic physical property of the earth's atmosphere, it cannot be measured or observed. In principle, an exact copy of the earth and the solar system would have to be created, then a small perturbation applied to the copy, and the future consequences observed. In studies like this paper, we are trying to estimate the intrinsic limit by simulating such quasi-identical copies, pretending that we know the initial atmospheric state exactly and assuming that the model provides a sufficiently accurate approximation of the atmosphere (perfect model assumption). Reanalyzes like ERA5 (or indeed any other analysis) however never come close enough to the true state to enable an observation of the butterfly effect, since our current observational and assimilation system still has very significant errors.

Hence, the neural network during its training can only learn an approximated development of the atmosphere within the range of our current assimilation uncertainties. The magnitudes of these uncertainties are represented by the 100% perturbations in the initial condition ensemble and therefore the AI-based models can only infer how such uncertainties would grow. Because of these limitations, we consider it very unlikely that other currently available AI-based models like FourCastNet (Pathak et al., 2022) or GraphCast (Lam et al., 2022), trained with similar data would produce significantly different results than the Pangu model tested here. However, models like GraphCast or the model used in Weyn et al. (2021) take two initial conditions as input, separated by a certain time interval (e.g., 6 hr). It would be interesting to see if such models could propagate the fast growth rates associated with the intrinsic limit further into the future, if they were present in the two initial conditions. But this would require a PDE-based model like ICON-HR-0.1% to generate the initial conditions and therefore essentially "inform" the neural network about the existence of the butterfly effect.

As stated in the introduction, the essence of the butterfly effect is upscale propagation of fast-growing uncertainties on small scales, mainly related to convection and precipitation. In coarser resolution models (both AI and PDE-based), the required small-scale variability is missing. In past research, we reintroduced the missing variability into coarse-resolution models by using a stochastic convection scheme, which led to faster error growth and a potentially more realistic simulation of the butterfly effect and a more accurate estimation of the intrinsic limit (Selz, 2019; Selz & Craig, 2015a; Selz et al., 2022). Similar methods could be used in AI-based models. Weyn et al. (2021) for example, created a set of (slightly) different models by randomizing the seed in the AI training process. This set of different models can be subsequently used to generate an ensemble of forecasts, analogous to stochastic parameterizations in current operational ensemble systems. Another approach could be to stochastically infer the missing small-scale variability by applying methods of super-resolution (Harris et al., 2022; Leinonen et al., 2020), in which ensemble skill metrics are part of the loss function. It will interesting to investigate if global prediction systems based on these techniques improve simulations of the butterfly effect and the characteristic amplitude-dependence of the error growth rate.

In any case, the failure of AI-based models to simulate the butterfly effect does not per se disqualify them from producing reliable ensembles in operational weather forecasting as it is unclear to what extent this ability is relevant for weather predictions at the current level of initial condition uncertainty (see introduction). Our previous study showed that upscale error growth processes from convection are currently unimportant on average, since they are overpowered by growth on synoptic scales from the relatively large initial uncertainties there (Selz et al., 2022). This type of error growth is simulated by the AI-based model quite well. But because this picture only represents the average mid-latitude conditions this may be different in certain extreme meteorological situations, as for example, described by Rodwell et al. (2013). It might also be different in the tropics, where convective processes are much stronger and more prominent, while large-scale processes are relatively weak and linear (Judt, 2020).

Data Availability Statement

The "Pangu Weather" model is available on GitHub (https://github.com/198808xc/Pangu-Weather). The ICON model code is restricted software and cannot be publicly shared. ECMWF analyses and EDA perturbations are available via the MARS-archive at ECMWF (https://apps.ecmwf.int/mars-catalogue/?class=od, restricted access). ERA5 reanalysis data is available via the Copernicus Climate Data Store (Hersbach et al., 2017). The output data from Pangu and ICON that has been evaluated in this paper can be retrieved from Selz and Craig (2023).

Acknowledgments
The research leading to this paper was carried out as part of the Collaborative Research Center SFB/TRR 165 "Waves to Weather" within project A1: "Upscale impact of diabatic processes from convective to near-hemispheric scale" funded by the German Research Foundation (DFG). The use of ECMWF's computing and archive facilities is gratefully acknowledged. We wish to thank Peter Dueben and an anonymous reviewer for their comments on the original manuscript. Open Access funding enabled and organized by Projekt DEAL.

References

Augier, P., & Lindborg, E. (2013). A new formulation of the spectral energy budget of the atmosphere, with application to two high-resolution general circulation models. *Journal of the Atmospheric Sciences*, *70*(7), 2293–2308. https://doi.org/10.1175/jas-d-12-0281.1

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), 1–6. https://doi.org/10.1038/s41586-023-06185-3

Craig, G. C., Fink, A. H., Hoose, C., Janjić, T., Knippertz, P., Laurian, A., et al. (2021). Waves to weather: Exploring the limits of predictability of weather. *Bulletin of the American Meteorological Society*, *102*(11), E2151–E2164. https://doi.org/10.1175/bams-d-20-0035.1

Durran, D. R., & Gingrich, M. (2014). Atmospheric predictability: Why butterflies are not of practical importance. *Journal of the Atmospheric Sciences*, *71*(7), 2476–2488. https://doi.org/10.1175/jas-d-14-0007.1

Harris, L., McRae, A. T., Chantry, M., Dueben, P. D., & Palmer, T. N. (2022). A generative deep learning approach to stochastic downscaling of precipitation forecasts. *Journal of Advances in Modeling Earth Systems*, *14*(10), e2022MS003120. https://doi.org/10.1029/2022ms003120

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

Isaksen, L., Bonavita, M., Buizza, R., Fisher, M., Haseler, J., Leutbecher, M., & Raynaud, L. (2010). Ensemble of data assimilations at ECMWF. ECMWF Technical Memoranda 636. Retrieved from https://www.ecmwf.int/sites/default/files/elibrary/2010/10125-ensemble-data-assimilations-ecmwf.pdf

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

Judt, F. (2020). Atmospheric predictability of the tropics, middle latitudes, and polar regions explored through global storm-resolving simulations. *Journal of the Atmospheric Sciences*, *77*(1), 257–276. https://doi.org/10.1175/jas-d-19-0116.1

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Pritzel, A., et al. (2022). Graphcast: Learning skillful medium-range global weather forecasting. arXiv preprint arXiv:2212.12794.

Leinonen, J., Nerini, D., & Berne, A. (2020). Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(9), 7211–7223. https://doi.org/10.1109/tgrs.2020.3032790

Lorenz, E. N. (1969). The predictability of a flow which possesses many scales of motion. *Tellus*, *21*(3), 289–307. https://doi.org/10.1111/j.2153-3490.1969.tb00444.x

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

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. arXiv preprint arXiv:2202.11214.

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

Rodwell, M. J., & Wernli, H. (2023). Uncertainty growth and forecast reliability during extratropical cyclogenesis. *Weather and Climate Dynamics*, *4*(3), 591–615. https://doi.org/10.5194/wcd-4-591-2023

Rotunno, R., & Snyder, C. (2008). A generalization of Lorenz's model for the predictability of flows with many scales of motion. *Journal of the Atmospheric Sciences*, *65*(3), 1063–1076. https://doi.org/10.1175/2007jas2449.1

Schulzweida, U. (2022). *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., & Craig, G. (2023). Data for "can artificial intelligence-based weather prediction models simulate the butterfly effect?" [Dataset]. LMU Munich, Faculty of Physics. https://doi.org/10.57970/e61hw-rrz34

Selz, T., & Craig, G. C. (2015a). Simulation of upscale error growth with a stochastic convection scheme. *Geophysical Research Letters*, *42*(8), 3056–3062. https://doi.org/10.1002/2015gl063525

Selz, T., & Craig, G. C. (2015b). 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., 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

Sun, Y. Q., & Zhang, F. (2016). Intrinsic versus practical limits of atmospheric predictability and the significance of the butterfly effect. *Journal of the Atmospheric Sciences*, *73*(3), 1419–1438. https://doi.org/10.1175/jas-d-15-0142.1

Tempest, K. I., Craig, G. C., & Brehmer, J. R. (2023). Convergence of forecast distributions in a 100,000-member idealised convective-scale ensemble. *Quarterly Journal of the Royal Meteorological Society*, *149*(752), 677–702. https://doi.org/10.1002/qj.4410

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

Weyn, J. A., Durran, D. R., Caruana, R., & Cresswell-Clay, N. (2021). Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *Journal of Advances in Modeling Earth Systems*, *13*(7), e2021MS002502. https://doi.org/10.1029/2021ms002502

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

Erratum

The originally published version of this article omitted the following statement at the end of the Acknowledgments: "Open Access funding enabled and organized by Projekt DEAL." The error has been corrected, and this may be considered the authoritative version of record.