On the effective resolution of AI weather prediction models

Tobias Selz[1], Wessel Bruinsma[2], George C. Craig[3], Stratis Markou[4], Richard Turner[4], and Anna Vaughan[5]

[1]Deutsches Zentrum fur Luft- und Raumfahrt DLR Abteilung Atmospharenprozessoren
[2]Microsoft Research
[3]Institute of Meteorology - University of Munich
[4]University of Cambridge
[5]Department of Earth Sciences, University of Cambridge, Cambridge, UK

March 08, 2025

Abstract

In this study, we investigate the effective resolution of deterministic AI weather prediction models. We find that an ideal, perfectly trained AI model follows the mean of the predictive distribution for the lead time interval which is used in its loss function during training. We demonstrate the consequences and limitations of this result with forecast data from various AI models, including Aurora, Pangu, GraphCast and GenCast and we compare them to ensemble and deterministic forecasts from the European Centre for Medium Range Weather Forecasting. We further demonstrate the impact of the resolution on mean-square error scores and suggest a method for a fairer comparison of two models with different effective resolution.

On the effective resolution of AI weather prediction models

T. Selz[1], W. P. Bruinsma[2], G. C. Craig[3], S. Markou[4], R. E. Turner[4,5], A. Vaughan[6]

[1]Deutsches Zentrum für Luft- und Raumfahrt, Oberpfaffenhofen, Germany
[2]Microsoft Research AI for Science, Amsterdam, Netherlands
[3]Ludwig-Maximilians-Universität, München, Germany
[4]Department of Engineering, University of Cambridge, Cambridge, UK
[5]The Alan Turing Institute, London, UK
[6]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK

Key Points:

- The effective resolution of an ideal AI model is determined by the spectrum of the ensemble mean at the lead times used in the loss function
- Real-world AI models approximate this behavior, but with a bias towards spatial smoothing
- Smooth models get better scores by avoiding the double-penalty effect

Corresponding author: Tobias Selz, `tobias.selz@lmu.de`

Abstract
In this study, we investigate the effective resolution of deterministic AI weather prediction models. We find that an ideal, perfectly trained AI model follows the mean of the predictive distribution for the lead time interval which is used in its loss function during training. We demonstrate the consequences and limitations of this result with forecast data from various AI models, including Aurora, Pangu, GraphCast and GenCast and we compare them to ensemble and deterministic forecasts from the European Centre for Medium Range Weather Forecasting. We further demonstrate the impact of the resolution on mean-square error scores and suggest a method for a fairer comparison of two models with different effective resolution.

Plain Language Summary

In recent years, models based on artificial intelligence (AI) have become equally good or even better at predicting the weather than standard models, which are based on solving physical equations. However, AI models often produce overly smooth forecasts, which lack relevant small-scale spatial structures. Here, we develop a mathematical argument to better understand this low "effective resolution" and investigate its applicability on recently developed AI models. It turns out that the lead time interval that is used during training plays a crucial role. Ironically, smooth forecasts can produce better scores by ignoring the small-scale structures and appear better than they actually are. We suggest a method to correct for this sometimes unwanted effect and get to a fairer comparison.

1 Introduction

Recently, several weather prediction models became available which use artificial intelligence (AI) to compute a deterministic forecast of the atmospheric state from an initial state (e.g., Bi et al., 2023; Lam et al., 2023; Bodnar et al., 2024). They have been trained on past atmospheric data and use mean square error (MSE) or mean absolute error (MAE) metrics to estimate their loss during training. These models have achieved similar or even better scores relative to "standard" numerical weather prediction models, which are based on solvers of the fluid equations, most notably the leading operational model — the Integrated Forecasting System (IFS) from ECMWF.

The spatial resolution of a weather model is defined as the size of its grid boxes. However, its "true" or "effective" resolution is usually much lower and is defined as the smallest spatial scale where atmospheric structures are reproduced with realistic amplitudes. The lower the effective resolution of a model, the smoother the forecast fields appear visually. While the effective resolution of standard weather models is mostly constant with lead time and adjusted with a diffusion scheme, it is less clear what determines the effective resolution of AI models, which can also significantly change with lead time. In fact, many AI models seem to suffer from excess smoothing and loss of energy at small scales (Ben Bouallègue et al., 2024; Selz & Craig, 2023).

For MSE or MAE metrics, it is well known that the optimal prediction is the mean or median, respectively, of the predictive distribution (Section 8.2 of Hsieh, 2023). Hence, one might expect that an AI forecast is closely related to the mean of an ensemble forecast. However, it is difficult to see such a relationship in practice (Bonavita, 2024).

The effective resolution of a weather prediction model is important for several reasons. First, the low computational cost of running AI models enables the creation of large ensembles to more accurately represent the forecast distribution. However, if each member has a low effective resolution or even resembles an ensemble mean, crucial variability will be missing. Second, extreme events are often caused by a superposition of features on many scales and a low resolution model may systematically underestimate them (e.g., Charlton-Perez et al., 2024). Third, for performance comparisons based on (root) mean square errors, smooth predictions will lead to better scores by avoiding the double-penalty effect, especially at long lead times (Ben Bouallègue et al., 2024; Bonavita, 2024), which has been framed as the "accuracy–activity trade-off" (Ben Bouallègue et al., 2024). Hence the question arises to what extent the better scores of the AI models are an artifact of their smoothness.

In this research letter, we show what effective resolution can be expected from the AI model in the ideal case of infinite capacity and perfect training and clarify the relationship between AI model predictions and the ensemble mean or median. Using forecasts from recent AI models, we then explore the practical validity of this argument and its limitations. Finally, we analyze and explain the effect of smoothing on error scores and suggest a spectral rescaling method for a "fairer", resolution-independent comparison.

2 Models, Data and Methods

2.1 Mathematical argument

We start by presenting a mathematical argument that connects the effective resolution of the AI model to the design of the loss function. Consider a true initial condition state vector x_{t_0}, from which we want to calculate a prediction $\hat{x}_t^\theta(x_{t_0})$ using an AI model, where t_0 and t refer to the forecast init and valid time, respectively, and θ to the set of learnable parameters of the model. Since the initial state is typically estimated with a certain amount of uncertainty which will grow with forecast lead time $\tau = t - t_0$, perfect forecasts from such imperfect initial states will be samples from a predictive distribution $p(x_t|x_{t_0})$.

With the training of an AI system, one tries to estimate the set of parameters θ^* which minimize the expectation of a distance metric between model forecasts $\hat{x}_t^\theta(x_{t_0})$ and true states x_t, the so-called loss function. Here, we assume a simple L2 metric over the normalized state vector and discuss other metrics below. In an ideal setting, the expectation of the loss function is taken over all possible initial and final states, hence

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\ \mathbb{E}_{p(x_t, x_{t_0})}\Big[||x_t - \hat{x}_t^\theta(x_{t_0})||^2\Big]. \tag{1}$$

With the law of total expectation and by expanding the square, this can be rewritten as

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\ \mathbb{E}_{p(x_{t_0})}\Big[||\mu_{t|t_0} - \hat{x}_t^\theta(x_{t_0})||^2\Big], \tag{2}$$

where we have defined the mean of the predictive distribution

$$\mu_{t|t_0} := \int \mathrm{d}x_t\ x_t\, p(x_t|x_{t_0}). \tag{3}$$

Consequently, the optimal prediction is the mean of the predictive distribution, i.e.:

$$\hat{x}_t^{\theta^*}(x_{t_0}) = \mu_{t|t_0}. \tag{4}$$

Some AI models use multiple time steps $(t_1, \ldots, t_n)$ inside the loss function and average over the individual loses:

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\ \mathbb{E}_{p(x_{t_n}, \ldots, x_{t_1}, x_{t_0})}\Big[\sum_{t'=t_1}^{t_n} ||x_{t'} - \hat{x}_{t'}^\theta(x_{t_0})||^2\Big]. \tag{5}$$

We will refer to this averaging period as the "lead time training interval"

$$\tau_{\text{train}} := t_n - t_0. \tag{6}$$

With the linearity of the expectation and the above we get

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{t'=t_1}^{t_n} \mathbb{E}_{p(x_{t_0})} \left[||\mu_{t'|t_0} - \hat{x}^{\theta}_{t'}(x_{t_0})||^2 \right]. \tag{7}$$

Hence an optimal prediction will follow the mean of the predictive distribution over τ_{train},

$$\hat{x}^{\theta^*}_t(x_{t_0}) = \mu_{t|t_0}, \quad \text{for } t \in t_0 + [\tau_1, \ldots, \tau_{\text{train}}]. \tag{8}$$

As we will see later in detail, this result has direct implications with respect to the effective resolution of the model, since unpredictable small-scale structures cancel out in the mean.

A similar result holds for other loss functions: In the case of the widely used L1 metric it can be shown that an ideal prediction will follow the median of the predictive distribution instead of the mean. Since most atmospheric variables have approximately symmetric predictive distributions, the mean and median are similar.

For real-world AI models the expectation in the ideal loss function needs to be replaced by averages over a training dataset,

$$L \sim \sum_{t_0} \sum_{\tau} \sum_{j} w_j \left(x^{(j)}_{t_0+\tau} - \hat{x}^{\theta\,(j)}_{t_0,\tau} \right)^2, \tag{9}$$

with j indexing the model state vector (grid box, level, variable). Mostly, ERA5 reanalysis (Hersbach et al., 2017) and IFS operational analysis have been used with initial times (t_0) from the satellite era (since 1979) as estimates of the truth. It is common to insert weighting factors w_j into the loss function (e.g., Bi et al., 2023). Also note that some AI models target differences rather than the variable values directly. However, none of these modifications affects the optimality results stated above.

Aside from these simple approaches, more complicated loss functions have sometimes been used, which also include non-linear functions of the state vector like spectra (e.g., Kochkov et al., 2024). In such cases the presented mathematical argument may not apply.

The ensemble median or mean is the target of training, but may not be achieved in practice. Neural networks appear to exhibit a spectral bias (Xu et al., 2019; Rahaman et al., 2019), where large spatial scales are learned first, and small scales may not be learned at all (Chattopadhyay et al., 2024). Therefore, we hypothesize that AI models due to lack of capacity or incomplete training will tend to be even smoother than the mean.

2.2 AI-model forecasts and data

To test the applicability of the mathematical argument, we analyze the effective resolution of several different AI models.

Aurora (Bodnar et al., 2024) is a transformer-based model. Its basic version, intended as a foundation model, is trained on a mixture of forecasts, analysis data, reanalysis data, and climate simulations. Here, we consider two versions with additional fine-tuning on IFS-HRES data. One version uses a short lead time training interval of only the first two time steps (6 h, 12 h), which we refer to as Aurora-S (for short). The other version uses a long lead time training interval of ten days, which we will call Aurora-L (for long).

Pangu (Bi et al., 2023) is also a transformer-based model, which was trained on ERA5 only. It comes in 4 different versions that perform forecasts for 4 different lead times (1 h, 3 h, 6 h, 24 h). The 1-h, 3-h, and 6-h models produce far less accurate forecasts than the 24-h model and are intended to be used only to successively fill in time steps. However, for the purpose of this study, we run each of these models individually. The lead time training interval for all of these models is only one time step.

GraphCast (Lam et al., 2023) is an AI model based on a graph neural network. Here we will not use the paper version, but the "operational" version, which has additional training on IFS-HRES data.

GenCast (Price et al., 2025), unlike the previous models, is trained to generate samples from the forecast distribution. It creates forecasts from denoising random fields. For the purpose of this paper, we only consider a single ensemble member. Like with GraphCast, we use the "operational" version, which in addition to ERA5 has been trained on IFS-HRES data.

All of these models use a regular lat-lon grid with 0.25° grid spacing and either a simple L1 or L2 metric in their loss function. With each model, we conducted a sample of 12 forecasts, initialized on the first day of each month of the year 2024. Unless stated otherwise, the presented results are averages over these cases to reduce random variability. All forecasts are carried out for 15 days lead time, except for Pangu-1h, which quickly became unstable. Regardless of its training dataset, we initialize every AI model with the IFS operational analysis.

To estimate the effective resolution of the models, we consider the kinetic energy spectrum at the upper troposphere (300 hPa), which follows known power laws (e.g., Nastrom & Gage, 1985). Kinetic energy spectra are computed based on global spherical harmonic coefficients of divergence (d) and vorticity (ζ), which are calculated from the horizontal wind using the Climate Data Operators (CDO; Schulzweida, 2024). The kinetic energy of a total wave number l is then given by (see e.g., Augier & Lindborg, 2013)

$$\mathrm{KE}(l) = \frac{r^2}{2l(l+1)} \sum_{m=-l}^{l} \left(|\zeta_{lm}|^2 + |d_{lm}|^2 \right), \tag{10}$$

where r is the radius of the earth and a wavelength $\lambda = 2\pi r/l$ is attributed to the global wave number l.

Finally, we need an estimate of the predictive distribution (3) to test the applicability of the mathematical argument. This will be taken from the ECMWF ensemble prediction system (IFS-ENS), a 50-member ensemble constructed from perturbations to sample uncertainty in the initial conditions and the model (see `https://www.ecmwf.int`). Here, we only show empirical results using the mean, since mean and median are similar for upper tropospheric winds but the median is more prone to sampling error.

The ensemble also includes an unperturbed control simulation (IFS-CTL), which since the resolution upgrade in June 2023 is identical to the former high-resolution deterministic run (HRES) and will be used as reference. For validation, the IFS operational analysis is used as the ground truth.

3 Results

3.1 Effective resolution and ensemble mean

We start by investigating the effective resolution of the Aurora-S and Aurora-L model, which differ greatly in their lead time training interval (12 hours versus 10 days), but are otherwise identical. Figure 1 shows their kinetic energy spectra for four different lead times. The IFS ensemble mean serves as estimator of the predictive distribution. Due to uncertainty growth from initial condition and model uncertainty, as the forecast lead time increases more and more spatial scales become unpredictable, which leads to their cancellation in the ensemble mean. This process starts at the smallest scales and successively affects larger and larger scales with increasing lead time (e.g., Selz et al., 2022).

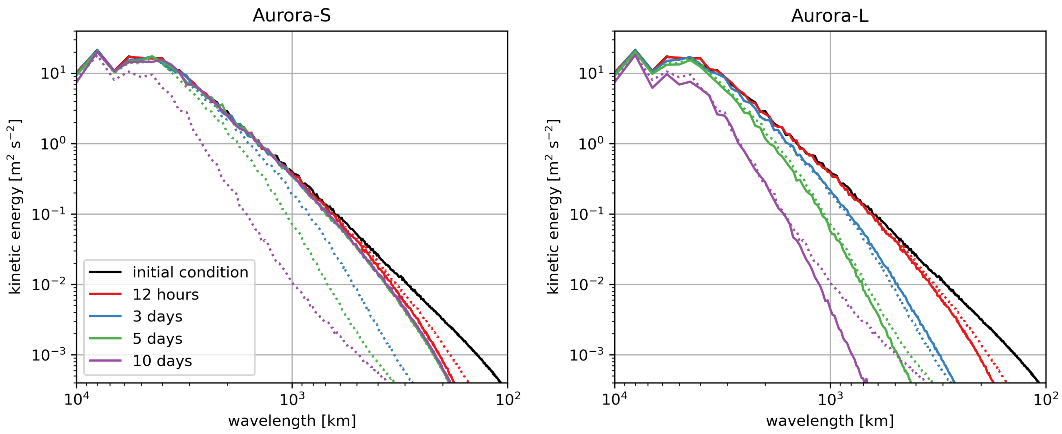

Figure 1. Kinetic energy spectra of Aurora-S (left) and Aurora-L (right), for several forecast lead times (solid lines). The dashed lines indicate the spectra of the IFS ensemble mean.

Hence, the "effective resolution" of the IFS ensemble mean continuously decreases with lead time and the kinetic energy becomes unrealistically low on larger and larger scales.

Looking at the Aurora-S simulations, the spectrum indicates an initial loss of small-scale energy in the first 12 hours, but stays approximately constant afterwards. For scales larger than about 300 km, the spectrum of Aurora-S stays close to the 12-h IFS ensemble mean. In contrast, the Aurora-L simulations constantly lose energy over lead time and follow the IFS ensemble mean closely, at least for amplitudes larger than 10^{-2} m^2 s^{-2}. The discrepancy below is due to sampling errors from the relatively small IFS ensemble. Also keep in mind that the IFS ensemble mean is only an estimate of the predictive distribution.

These results clearly illustrate the importance of the lead time training interval for the effective resolution of deterministic AI models. While Aurora-S produces a largely stable spectrum, Aurora-L suffers from a continuous loss of kinetic energy and effective resolution and closely follows the IFS ensemble mean. To further demonstrate the significance of these differences, Fig. 2 shows maps from a single 10-day forecast from both Aurora models, the IFS-CTL and the IFS ensemble mean. Aurora-S and the IFS-CTL show pronounced Rossby wave structures with troughs and ridges and associated meridional winds. Although different from each other and from the truth, both states are approximate realizations of the atmospheric flow or samples from the predictive distribution. On the other hand, the loss of small-scale kinetic energy of the Aurora-L forecasts results in highly smoothed spatial fields with strongly damped Rossby waves. The resemblance of Aurora-L to the IFS-ensemble mean is clearly visible. These forecasts are not possible realizations of the atmospheric flow, but they estimate the expectation of the predictive distribution.

3.2 Kinetic energy time series

In order to test the effective resolution and the applicability of the mathematical argument on further AI models, we integrate the kinetic energy between scales of 400 km and 4000 km. This results in a time series for each model that quantifies kinetic energy loss, which is shown in Figure 3, also including the IFS ensemble mean as reference.

We start with discussing the four different versions of Pangu, where the lead time training interval is only the first time step, i.e., 1 h, 3 h, 6 h, and 24 h, respectively. The kinetic energies at the end of the training intervals are close to the IFS ensemble mean,

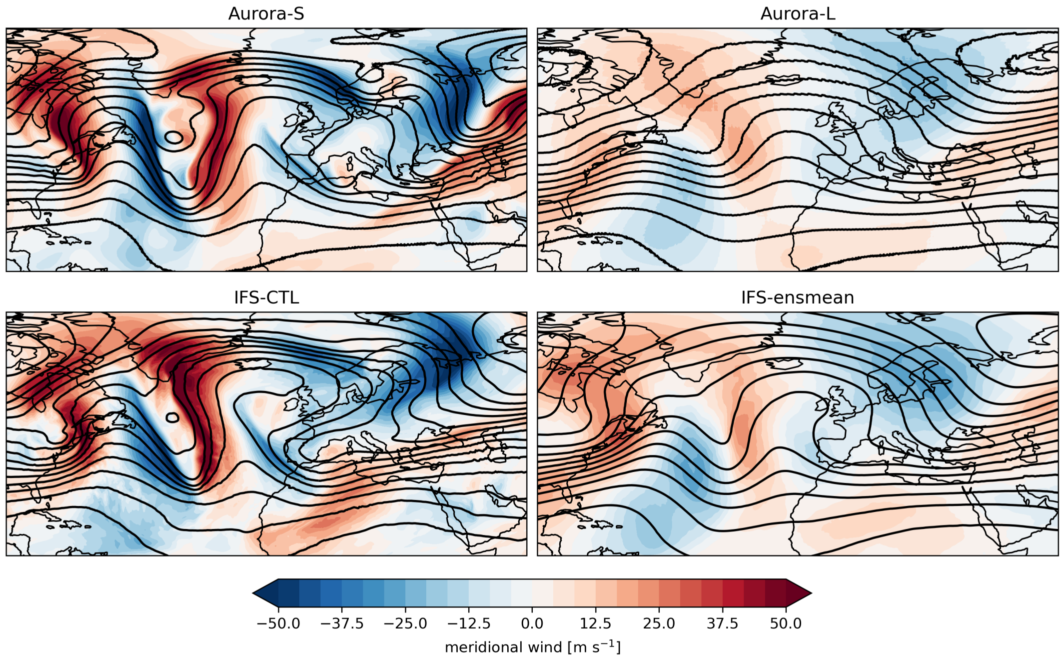

Figure 2. 10-day forecasts of 300 hPa meridional wind (color) and geopotential (lines, spacing $1000\,\mathrm{m}^2\,\mathrm{s}^{-2}$) over the North Atlantic and Europe for four different experiments. The forecasts were started on 1 Jan 2024, 0 UTC.

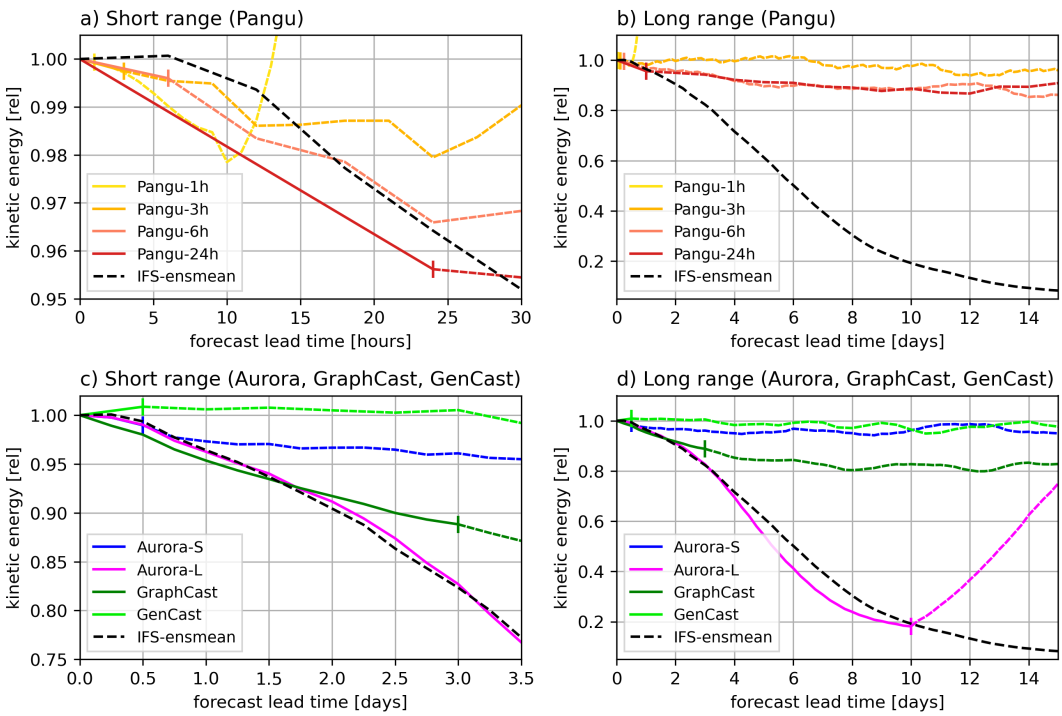

Figure 3. 300 hPa kinetic energy between 400 km and 4000 km wavelength over lead time, relative to initial condition. The plots on the left show a zoom into the initial period. Top and bottom rows show different sets of models. Solid lines indicate lead times within the training interval ($\tau \leq \tau_{\mathrm{train}}$), and dashed lines indicate later lead times. A vertical bar is marking τ_{train}.

but slightly too low. Most notably, the 24-h model at its first time step has a much lower resolution compared to the other three models, which are roughly similar. After the training interval, the 3-h, 6-h, and 24-h model further lose some kinetic energy, but after a few days show a more stable spectrum. The 1-h model however, after an initial loss of kinetic energy, quickly becomes unstable.

For the two Aurora models, Fig. 3 confirms the findings already discussed above: Aurora-S creates a basically stable spectrum, slightly below the IFS-ensemble mean value at the end of the 12-h training interval, while Aurora-L produces a constantly decaying spectrum, closely following the IFS ensemble mean over the 10-day training interval. Note however, that the kinetic energy of Aurora-L is increasing again after this 10-day period, which points to an accumulation of unphysical noise and indicates an unstable model that is not suitable for longer forecasts.

The GraphCast model with its 3-day training interval only roughly follows the IFS ensemble mean, being slightly smoother for the first 1.5 days, and less smooth for the second 1.5 days. This latter behavior contradicts our expectations by producing a forecast with higher effective resolution than the ensemble mean. However, GraphCast was trained using a curriculum approach in which training stated with a single time interval and then slowly increased the lead time interval out to three days. This combined with the fact that GraphCast is a relatively small model is likely lead to the behavior noted above. After the 3 days there is some further decay of kinetic energy, but the spectrum remains stable after about 6-7 days.

GenCast, which is not trained to approximate the ensemble mean or median, but to generate samples from the full distribution, is best able to retain the initial spectrum at all lead times.

3.3 Impact of the resolution on mean-square error scores

A standard way to evaluate the quality of deterministic weather forecasts is to compute the spatially averaged squared difference of some variable to a representation of the truth, referred to as mean-square error. Among others, Ben Bouallègue et al. (2024) demonstrated, that smooth ("low activity") forecasts can lead to better MSE scores by avoiding the double-penalty effect. With the help of the kinetic energy spectrum, we formally explain the reason for the double-penalty effect and confirm it with our simulation data.

An area-weighted mean-square error over the entire globe can equally be computed from spherical harmonics expansions, since Parseval's identity applies. This allows for a scale-dependent formulation of the error, which for error kinetic energy (EKE) reads

$$\mathrm{EKE}(l) = \frac{r^2}{2l(l+1)} \sum_{m=-l}^{l} \Big(|\hat{\zeta}_{lm} - \zeta_{lm}|^2 + |\hat{d}_{lm} - d_{lm}|^2 \Big), \tag{11}$$

where the hat indicates the forecast and non-hat symbols indicate the truth (a similar formalism can be applied to limited domains using Fourier or Cosine transforms). The scale-dependent EKE of the 10-day forecasts is plotted in Fig. 4a, normalized with the kinetic energy (10) of the analysis. For reference, the equally normalized kinetic energy spectrum is shown in Fig. 4b.

To interpret these plots and to understand the double-penalty effect, we expand the absolute square difference,

$$\sum_m |\hat{\zeta}_{lm} - \zeta_{lm}|^2 = \sum_m \Big[(\hat{r}_{lm} - r_{lm})^2 + 2\hat{r}_{lm} r_{lm} \big(1 - \cos(\hat{\phi}_{lm} - \phi_{lm})\big) \Big], \tag{12}$$

where r_{lm} and ϕ_{lm} are amplitude and phase of the complex number ζ_{lm}, respectively. A similar expression holds for any other variable. Consider a mode l, that is no longer

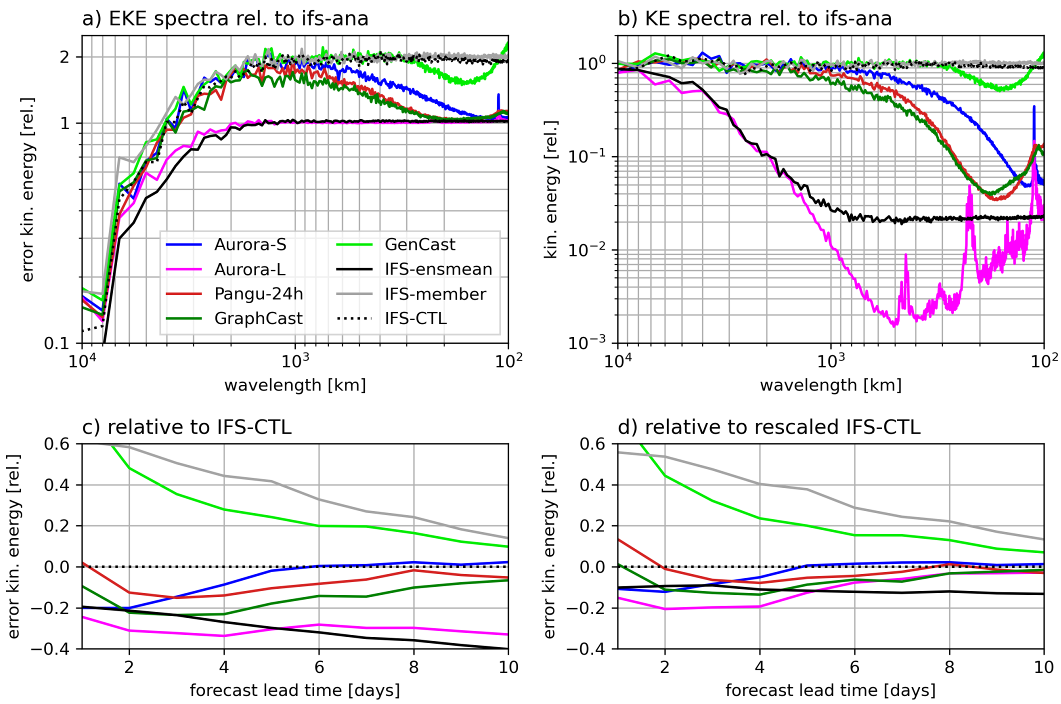

Figure 4. (a) Error kinetic energy spectra of 10-day forecasts over wavelength, relative to the kinetic energy spectrum of the IFS analysis. (b) Same as a, but for kinetic energy spectra. (c) Globally averaged EKE relative to IFS-CTL, computed using (11) and summing over l. (d) Same as c, but relative to a rescaled version of the IFS-CTL by applying (13). Note that these rescale factors differ, depending on the model IFS-CTL was compared to.

predictable. If the model returns zero for that mode, the second term on the left hand side in (12) vanishes and the error equals the amplitude of the analysis spectrum. On the other hand, if the model maintains the correct amplitude but predicts a random phase, the first term vanishes and the error equals *twice* the analysis spectrum (since the expectation of the cosine term is zero) and therefore twice the error compared to predicting zeros (hence double-penalty).

This relation between the error (EKE) and the amplitude (KE) for unpredictable modes becomes evident from our data by comparing Figs. 4a and b: Aurora-L and the IFS ensemble mean produce a normalized EKE of one for scales smaller than 2000 km and at the same time an amplitude close to zero. The other models resemble the IFS-CTL for scales larger than around 1000 km, producing an EKE of two, but an almost realistic amplitude. Towards small scales, the normalized EKE of all AI models except GenCast drops to one as a consequence of their decaying KE. The consequence of the double-penalty effect can also clearly be seen in the EKE time series (Fig. 4c), where smooth forecasts (IFS ensemble mean and Aurora-L) clearly outperform the IFS-CTL and every other model, most significantly at long lead times.

As demonstrated, the scores of the AI models are enhanced by the cancellation of unpredictable modes, which does not indicate a “true” advantage. But the question remains, to what extent? One possibility to exclude the smoothing benefit from a comparison of two models is to equalize their spectra before calculating the EKE or any other mean square error. This can be done by rescaling (damping) the spectral modes of model B to the amplitude of the smoother model A, i.e.,

$$\zeta^{\mathrm{B}}_{lm} \longrightarrow \sqrt{\frac{\sum_m |\zeta^{\mathrm{A}}_{lm}|^2}{\sum_m |\zeta^{\mathrm{B}}_{lm}|^2}}\, \zeta^{\mathrm{B}}_{lm}, \tag{13}$$

and similarly for other variables.

The result of such a comparison is shown in Fig. 4d, where the IFS-CTL spectrum was rescaled to the AI model spectrum. One can see, that the superior skills of the IFS ensemble mean and Aurora-L from Fig. 4c are greatly reduced, especially at long lead times. Indeed for lead times greater than about one week, all AI models perform equally well compared to IFS-CTL, or rather equally badly since there is little practical predictability remaining (Buizza & Leutbecher, 2015; Selz et al., 2022). The difference between Figs. 4c and d is directly correlated to the amount of smoothing produced by the models: It is large for the IFS ensemble mean and Aurora-L, but small for models that approximately maintain the KE spectrum, like Aurora-S, Pangu and GenCast. Note that GenCast is trained to generate samples of the predictive distribution and hence introduces perturbations, which lead to larger errors, especially at early lead times. An even slightly worse degradation of the EKE can be seen from an individual member of the IFS ensemble.

4 Discussion

In summary, we demonstrated with a mathematical argument that the lead time interval in the loss function crucially determines the kinetic energy spectrum and hence the effective resolution of an AI model. If perfectly trained, a model would follow the spectrum of an ideal ensemble mean over that interval and continuously drop unpredictable modes, leading to increasingly smooth forecasts. We also confirmed that smooth forecasts produce much better mean-square error scores by avoiding the double penalty effect and we suggested a method to correct for that.

From our findings, we can identify two basic approaches to weather forecasting with AI: Either a model could be designed to generate samples from the predictive distribution, in which case the lead time training interval should be kept as short as possible. Alternatively, a model could be designed to generate the expectation (the ensemble mean) of the predictive distribution, in which case the lead time training interval should extend to the entire intended forecast lead time.

Both approaches have their justification, however, they should not be mixed and it should be made clear, which one was chosen, since this has consequences for the usage of the model. Models of the first type (Aurora-S, Pangu) can be used to sample the forecast distribution by means of an ensemble, stated from an initial condition sample or using intrinsic stochastisity (GenCast). Each simulation resembles a possible state of the atmosphere that, at least approximately, is physically consistent. Models of the second type on the other hand (like Aurora-L) are not suitable to generate ensembles, do not produce possible realizations of the atmospheric flow and their output is physically inconsistent. However, they do resemble the remaining predictable structures in a single run and predictability can be inferred from the remaining spatial scales.

Although the lead time training interval is crucial for the model's effective resolution, it cannot explain every aspect of it. Most importantly, the presented mathematical argument does not hold for predictions outside of the lead time training interval. In this case, previous forecasts are being fed into the model, which resemble the mean and are much smoother than the training dataset. The reaction of the model to this inconsistency is largely unconstrained. In fact, some models in our study show instability after the lead time training interval and most models continue to lose some kinetic energy. In addition, no model was able to maintain the kinetic energy spectrum on scales smaller than about 300-400 km. Potential reasons for these effects include insufficient training, insufficient capacity, limited sample size of the training data or limitations in the design of the network.

Open Research Section

The AI-model weights, example code and documentation can be found on github: `https://github.com/google-deepmind/GraphCast`, `https://github.com/microsoft/aurora`, `https://github.com/198808xc/Pangu-Weather`. The spherical harmonic coefficients of the forecast data are available at `https://opendata.physik.lmu.de/H66gKyhITQ7qS51` (permanent link after acceptance). The IFS operational analyses, the IFS-CTL and IFS-ENS forecast were retrieved from ECMWF's operational archive (`https://apps.ecmwf.int/archive-catalogue/?class=od`). To obtain access, visit `https://www.ecmwf.int/en/forecasts/accessing-forecasts` for further information.

Acknowledgments

Richard E. Turner is supported by the EPSRC Probabilistic AI Hub and an EPSRC Prosperity Partnership with Microsoft Research. Stratis Markou acknowledges funding from the Vice Chancellor's and George & Marie Vergottis scholarship and the Qualcomm Innovation Fellowship. Anna Vaughan is supported by the UKRI Centre for Doctoral Training in Application of Artificial Intelligence to the study of Environmental Risks (AI4ER). The use of ECMWF's computing and archive facilities is gratefully acknowledged.

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.

Ben Bouallègue, Z., Clare, M. C., Magnusson, L., Gascon, E., Maier-Gerber, M., Janoušek, M., … others (2024). The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an operational-like context. *Bulletin of the American Meteorological Society*, *105*(6), E864–E883.

Ben Bouallègue, Z., et al. (2024). Accuracy versus activity. *AIFS Blog*. doi: 10.21957/8b50609a0f

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.

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Brandstetter, J., Garvan, P., … others (2024). Aurora: A foundation model of the atmosphere. *arXiv preprint arXiv:2405.13063*.

Bonavita, M. (2024). On some limitations of current machine learning weather prediction models. *Geophysical Research Letters*, *51*(12), e2023GL107377.

Buizza, R., & Leutbecher, M. (2015). The forecast skill horizon. *Quarterly Journal of the Royal Meteorological Society*, *141*(693), 3366–3382.

Charlton-Perez, A. J., Dacre, H. F., Driscoll, S., Gray, S. L., Harvey, B., Harvey, N. J., … others (2024). Do ai models produce better weather forecasts than physics-based models? a quantitative evaluation case study of storm ciarán. *npj Climate and Atmospheric Science*, *7*(1), 93.

Chattopadhyay, A., Sun, Y. Q., & Hassanzadeh, P. (2024). Challenges of learning multi-scale dynamics with ai weather models: Implications for stability and one solution. Retrieved from `https://arxiv.org/abs/2304.07029`

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., … Thépaut, J.-N. (2017). Complete era5 from 1940: Fifth generation of ecmwf atmospheric reanalyses of the global climate [dataset]. *Copernicus Climate Change Service (C3S) Data Store (CDS).* (Accessed on 12-Jul-2023) doi: 10.24381/cds.143582cf

Hsieh, W. W. (2023). *Introduction to environmental data science.* Cambridge University Press.

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., … others (2024). Neural general circulation models for weather and climate. *Nature*, *632*(8027), 1060–1066.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., … others (2023). Learning skillful medium-range global weather forecasting. *Science*, *382*(6677), 1416–1421.

Nastrom, G., & Gage, K. (1985). A climatology of atmospheric wavenumber spectra of wind and temperature observed by commercial aircraft. *J. Atmos. Sci*, *42*, 950–960.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., … others (2025). Probabilistic weather forecasting with machine learning. *Nature*, *637*(8044), 84–90.

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F. A., … Courville, A. (2019). *On the spectral bias of neural networks.* Retrieved from `https://arxiv.org/abs/1806.08734`

Schulzweida, U. (2024). *Cdo user guide.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.7112925` doi: 10.5281/zenodo.7112925

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

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.

Xu, L., Wang, S., & Tang, R. (2019). Probabilistic load forecasting for buildings considering weather forecasting uncertainty and uncertain peak load. *Applied energy*, *237*, 180–195.

On the effective resolution of AI weather prediction models

T. Selz[1], W. P. Bruinsma[2], G. C. Craig[3], S. Markou[4], R. E. Turner[4,5], A. Vaughan[6]

[1]Deutsches Zentrum für Luft- und Raumfahrt, Oberpfaffenhofen, Germany
[2]Microsoft Research AI for Science, Amsterdam, Netherlands
[3]Ludwig-Maximilians-Universität, München, Germany
[4]Department of Engineering, University of Cambridge, Cambridge, UK
[5]The Alan Turing Institute, London, UK
[6]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK

Key Points:

- The effective resolution of an ideal AI model is determined by the spectrum of the ensemble mean at the lead times used in the loss function
- Real-world AI models approximate this behavior, but with a bias towards spatial smoothing
- Smooth models get better scores by avoiding the double-penalty effect

Corresponding author: Tobias Selz, `tobias.selz@lmu.de`

Abstract

In this study, we investigate the effective resolution of deterministic AI weather prediction models. We find that an ideal, perfectly trained AI model follows the mean of the predictive distribution for the lead time interval which is used in its loss function during training. We demonstrate the consequences and limitations of this result with forecast data from various AI models, including Aurora, Pangu, GraphCast and GenCast and we compare them to ensemble and deterministic forecasts from the European Centre for Medium Range Weather Forecasting. We further demonstrate the impact of the resolution on mean-square error scores and suggest a method for a fairer comparison of two models with different effective resolution.

Plain Language Summary

In recent years, models based on artificial intelligence (AI) have become equally good or even better at predicting the weather than standard models, which are based on solving physical equations. However, AI models often produce overly smooth forecasts, which lack relevant small-scale spatial structures. Here, we develop a mathematical argument to better understand this low "effective resolution" and investigate its applicability on recently developed AI models. It turns out that the lead time interval that is used during training plays a crucial role. Ironically, smooth forecasts can produce better scores by ignoring the small-scale structures and appear better than they actually are. We suggest a method to correct for this sometimes unwanted effect and get to a fairer comparison.

1 Introduction

Recently, several weather prediction models became available which use artificial intelligence (AI) to compute a deterministic forecast of the atmospheric state from an initial state (e.g., Bi et al., 2023; Lam et al., 2023; Bodnar et al., 2024). They have been trained on past atmospheric data and use mean square error (MSE) or mean absolute error (MAE) metrics to estimate their loss during training. These models have achieved similar or even better scores relative to "standard" numerical weather prediction models, which are based on solvers of the fluid equations, most notably the leading operational model — the Integrated Forecasting System (IFS) from ECMWF.

The spatial resolution of a weather model is defined as the size of its grid boxes. However, its "true" or "effective" resolution is usually much lower and is defined as the smallest spatial scale where atmospheric structures are reproduced with realistic amplitudes. The lower the effective resolution of a model, the smoother the forecast fields appear visually. While the effective resolution of standard weather models is mostly constant with lead time and adjusted with a diffusion scheme, it is less clear what determines the effective resolution of AI models, which can also significantly change with lead time. In fact, many AI models seem to suffer from excess smoothing and loss of energy at small scales (Ben Bouallègue et al., 2024; Selz & Craig, 2023).

For MSE or MAE metrics, it is well known that the optimal prediction is the mean or median, respectively, of the predictive distribution (Section 8.2 of Hsieh, 2023). Hence, one might expect that an AI forecast is closely related to the mean of an ensemble forecast. However, it is difficult to see such a relationship in practice (Bonavita, 2024).

The effective resolution of a weather prediction model is important for several reasons. First, the low computational cost of running AI models enables the creation of large ensembles to more accurately represent the forecast distribution. However, if each member has a low effective resolution or even resembles an ensemble mean, crucial variability will be missing. Second, extreme events are often caused by a superposition of features on many scales and a low resolution model may systematically underestimate them (e.g., Charlton-Perez et al., 2024). Third, for performance comparisons based on (root) mean square errors, smooth predictions will lead to better scores by avoiding the double-penalty effect, especially at long lead times (Ben Bouallègue et al., 2024; Bonavita, 2024), which has been framed as the "accuracy–activity trade-off" (Ben Bouallègue et al., 2024). Hence the question arises to what extent the better scores of the AI models are an artifact of their smoothness.

In this research letter, we show what effective resolution can be expected from the AI model in the ideal case of infinite capacity and perfect training and clarify the relationship between AI model predictions and the ensemble mean or median. Using forecasts from recent AI models, we then explore the practical validity of this argument and its limitations. Finally, we analyze and explain the effect of smoothing on error scores and suggest a spectral rescaling method for a "fairer", resolution-independent comparison.

2 Models, Data and Methods

2.1 Mathematical argument

We start by presenting a mathematical argument that connects the effective resolution of the AI model to the design of the loss function. Consider a true initial condition state vector x_{t_0}, from which we want to calculate a prediction $\hat{x}_t^\theta(x_{t_0})$ using an AI model, where t_0 and t refer to the forecast init and valid time, respectively, and θ to the set of learnable parameters of the model. Since the initial state is typically estimated with a certain amount of uncertainty which will grow with forecast lead time $\tau = t - t_0$, perfect forecasts from such imperfect initial states will be samples from a predictive distribution $p(x_t|x_{t_0})$.

With the training of an AI system, one tries to estimate the set of parameters θ^* which minimize the expectation of a distance metric between model forecasts $\hat{x}_t^\theta(x_{t_0})$ and true states x_t, the so-called loss function. Here, we assume a simple L2 metric over the normalized state vector and discuss other metrics below. In an ideal setting, the expectation of the loss function is taken over all possible initial and final states, hence

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \; \mathbb{E}_{p(x_t, x_{t_0})} \Big[||x_t - \hat{x}_t^\theta(x_{t_0})||^2 \Big]. \tag{1}$$

With the law of total expectation and by expanding the square, this can be rewritten as

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \; \mathbb{E}_{p(x_{t_0})} \Big[||\mu_{t|t_0} - \hat{x}_t^\theta(x_{t_0})||^2 \Big], \tag{2}$$

where we have defined the mean of the predictive distribution

$$\mu_{t|t_0} := \int \mathrm{d}x_t \; x_t \, p(x_t|x_{t_0}). \tag{3}$$

Consequently, the optimal prediction is the mean of the predictive distribution, i.e.:

$$\hat{x}_t^{\theta^*}(x_{t_0}) = \mu_{t|t_0}. \tag{4}$$

Some AI models use multiple time steps $(t_1, \ldots, t_n)$ inside the loss function and average over the individual loses:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \; \mathbb{E}_{p(x_{t_n}, \ldots, x_{t_1}, x_{t_0})} \Big[\sum_{t'=t_1}^{t_n} ||x_{t'} - \hat{x}_{t'}^\theta(x_{t_0})||^2 \Big]. \tag{5}$$

We will refer to this averaging period as the "lead time training interval"

$$\tau_{\text{train}} := t_n - t_0. \tag{6}$$

With the linearity of the expectation and the above we get

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{t'=t_1}^{t_n} \mathbb{E}_{p(x_{t_0})} \Big[||\mu_{t'|t_0} - \hat{x}^{\theta}_{t'}(x_{t_0})||^2 \Big]. \tag{7}$$

Hence an optimal prediction will follow the mean of the predictive distribution over τ_{train},

$$\hat{x}^{\theta^*}_t(x_{t_0}) = \mu_{t|t_0}, \quad \text{for } t \in t_0 + [\tau_1, \ldots, \tau_{\text{train}}]. \tag{8}$$

As we will see later in detail, this result has direct implications with respect to the effective resolution of the model, since unpredictable small-scale structures cancel out in the mean.

A similar result holds for other loss functions: In the case of the widely used L1 metric it can be shown that an ideal prediction will follow the median of the predictive distribution instead of the mean. Since most atmospheric variables have approximately symmetric predictive distributions, the mean and median are similar.

For real-world AI models the expectation in the ideal loss function needs to be replaced by averages over a training dataset,

$$L \sim \sum_{t_0} \sum_{\tau} \sum_{j} w_j \left(x^{(j)}_{t_0+\tau} - \hat{x}^{\theta\,(j)}_{t_0,\tau} \right)^2, \tag{9}$$

with j indexing the model state vector (grid box, level, variable). Mostly, ERA5 reanalysis (Hersbach et al., 2017) and IFS operational analysis have been used with initial times (t_0) from the satellite era (since 1979) as estimates of the truth. It is common to insert weighting factors w_j into the loss function (e.g., Bi et al., 2023). Also note that some AI models target differences rather than the variable values directly. However, none of these modifications affects the optimality results stated above.

Aside from these simple approaches, more complicated loss functions have sometimes been used, which also include non-linear functions of the state vector like spectra (e.g., Kochkov et al., 2024). In such cases the presented mathematical argument may not apply.

The ensemble median or mean is the target of training, but may not be achieved in practice. Neural networks appear to exhibit a spectral bias (Xu et al., 2019; Rahaman et al., 2019), where large spatial scales are learned first, and small scales may not be learned at all (Chattopadhyay et al., 2024). Therefore, we hypothesize that AI models due to lack of capacity or incomplete training will tend to be even smoother than the mean.

2.2 AI-model forecasts and data

To test the applicability of the mathematical argument, we analyze the effective resolution of several different AI models.

Aurora (Bodnar et al., 2024) is a transformer-based model. Its basic version, intended as a foundation model, is trained on a mixture of forecasts, analysis data, reanalysis data, and climate simulations. Here, we consider two versions with additional fine-tuning on IFS-HRES data. One version uses a short lead time training interval of only the first two time steps (6 h, 12 h), which we refer to as Aurora-S (for short). The other version uses a long lead time training interval of ten days, which we will call Aurora-L (for long).

Pangu (Bi et al., 2023) is also a transformer-based model, which was trained on ERA5 only. It comes in 4 different versions that perform forecasts for 4 different lead times (1 h, 3 h, 6 h, 24 h). The 1-h, 3-h, and 6-h models produce far less accurate forecasts than the 24-h model and are intended to be used only to successively fill in time steps. However, for the purpose of this study, we run each of these models individually. The lead time training interval for all of these models is only one time step.

GraphCast (Lam et al., 2023) is an AI model based on a graph neural network. Here we will not use the paper version, but the "operational" version, which has additional training on IFS-HRES data.

GenCast (Price et al., 2025), unlike the previous models, is trained to generate samples from the forecast distribution. It creates forecasts from denoising random fields. For the purpose of this paper, we only consider a single ensemble member. Like with GraphCast, we use the "operational" version, which in addition to ERA5 has been trained on IFS-HRES data.

All of these models use a regular lat-lon grid with 0.25° grid spacing and either a simple L1 or L2 metric in their loss function. With each model, we conducted a sample of 12 forecasts, initialized on the first day of each month of the year 2024. Unless stated otherwise, the presented results are averages over these cases to reduce random variability. All forecasts are carried out for 15 days lead time, except for Pangu-1h, which quickly became unstable. Regardless of its training dataset, we initialize every AI model with the IFS operational analysis.

To estimate the effective resolution of the models, we consider the kinetic energy spectrum at the upper troposphere (300 hPa), which follows known power laws (e.g., Nastrom & Gage, 1985). Kinetic energy spectra are computed based on global spherical harmonic coefficients of divergence (d) and vorticity (ζ), which are calculated from the horizontal wind using the Climate Data Operators (CDO; Schulzweida, 2024). The kinetic energy of a total wave number l is then given by (see e.g., Augier & Lindborg, 2013)

$$\mathrm{KE}(l) = \frac{r^2}{2l(l+1)} \sum_{m=-l}^{l} \left(|\zeta_{lm}|^2 + |d_{lm}|^2 \right), \tag{10}$$

where r is the radius of the earth and a wavelength $\lambda = 2\pi r/l$ is attributed to the global wave number l.

Finally, we need an estimate of the predictive distribution (3) to test the applicability of the mathematical argument. This will be taken from the ECMWF ensemble prediction system (IFS-ENS), a 50-member ensemble constructed from perturbations to sample uncertainty in the initial conditions and the model (see `https://www.ecmwf.int`). Here, we only show empirical results using the mean, since mean and median are similar for upper tropospheric winds but the median is more prone to sampling error.

The ensemble also includes an unperturbed control simulation (IFS-CTL), which since the resolution upgrade in June 2023 is identical to the former high-resolution deterministic run (HRES) and will be used as reference. For validation, the IFS operational analysis is used as the ground truth.

3 Results

3.1 Effective resolution and ensemble mean

We start by investigating the effective resolution of the Aurora-S and Aurora-L model, which differ greatly in their lead time training interval (12 hours versus 10 days), but are otherwise identical. Figure 1 shows their kinetic energy spectra for four different lead times. The IFS ensemble mean serves as estimator of the predictive distribution. Due to uncertainty growth from initial condition and model uncertainty, as the forecast lead time increases more and more spatial scales become unpredictable, which leads to their cancellation in the ensemble mean. This process starts at the smallest scales and successively affects larger and larger scales with increasing lead time (e.g., Selz et al., 2022).

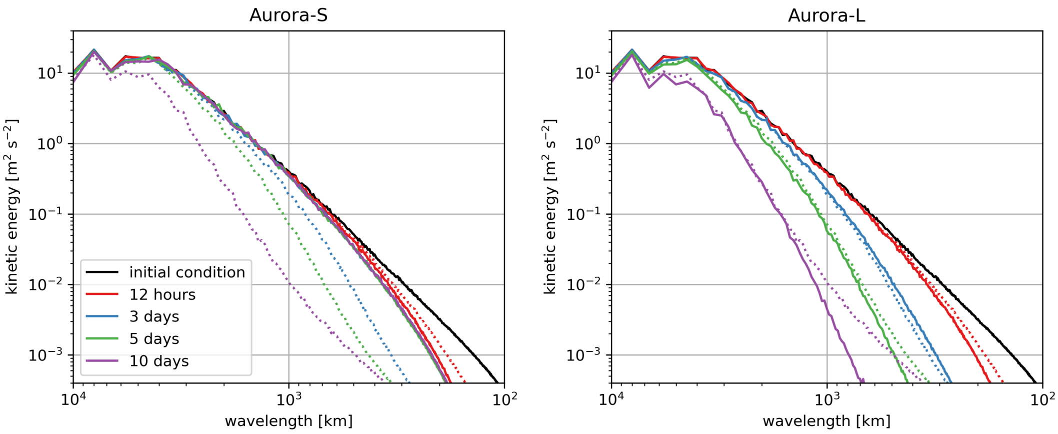

Figure 1. Kinetic energy spectra of Aurora-S (left) and Aurora-L (right), for several forecast lead times (solid lines). The dashed lines indicate the spectra of the IFS ensemble mean.

Hence, the "effective resolution" of the IFS ensemble mean continuously decreases with lead time and the kinetic energy becomes unrealistically low on larger and larger scales.

Looking at the Aurora-S simulations, the spectrum indicates an initial loss of small-scale energy in the first 12 hours, but stays approximately constant afterwards. For scales larger than about 300 km, the spectrum of Aurora-S stays close to the 12-h IFS ensemble mean. In contrast, the Aurora-L simulations constantly lose energy over lead time and follow the IFS ensemble mean closely, at least for amplitudes larger than 10^{-2} m^2 s^{-2}. The discrepancy below is due to sampling errors from the relatively small IFS ensemble. Also keep in mind that the IFS ensemble mean is only an estimate of the predictive distribution.

These results clearly illustrate the importance of the lead time training interval for the effective resolution of deterministic AI models. While Aurora-S produces a largely stable spectrum, Aurora-L suffers from a continuous loss of kinetic energy and effective resolution and closely follows the IFS ensemble mean. To further demonstrate the significance of these differences, Fig. 2 shows maps from a single 10-day forecast from both Aurora models, the IFS-CTL and the IFS ensemble mean. Aurora-S and the IFS-CTL show pronounced Rossby wave structures with troughs and ridges and associated meridional winds. Although different from each other and from the truth, both states are approximate realizations of the atmospheric flow or samples from the predictive distribution. On the other hand, the loss of small-scale kinetic energy of the Aurora-L forecasts results in highly smoothed spatial fields with strongly damped Rossby waves. The resemblance of Aurora-L to the IFS-ensemble mean is clearly visible. These forecasts are not possible realizations of the atmospheric flow, but they estimate the expectation of the predictive distribution.

3.2 Kinetic energy time series

In order to test the effective resolution and the applicability of the mathematical argument on further AI models, we integrate the kinetic energy between scales of 400 km and 4000 km. This results in a time series for each model that quantifies kinetic energy loss, which is shown in Figure 3, also including the IFS ensemble mean as reference.

We start with discussing the four different versions of Pangu, where the lead time training interval is only the first time step, i.e., 1 h, 3 h, 6 h, and 24 h, respectively. The kinetic energies at the end of the training intervals are close to the IFS ensemble mean,

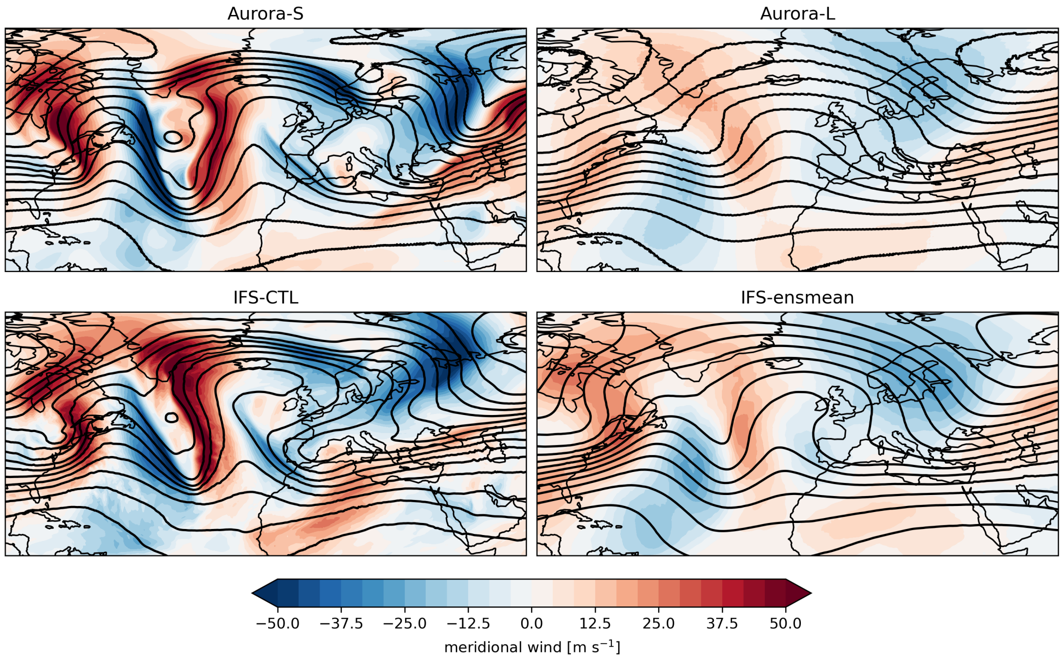

Figure 2. 10-day forecasts of 300 hPa meridional wind (color) and geopotential (lines, spacing $1000\,\mathrm{m}^2\,\mathrm{s}^{-2}$) over the North Atlantic and Europe for four different experiments. The forecasts were started on 1 Jan 2024, 0 UTC.

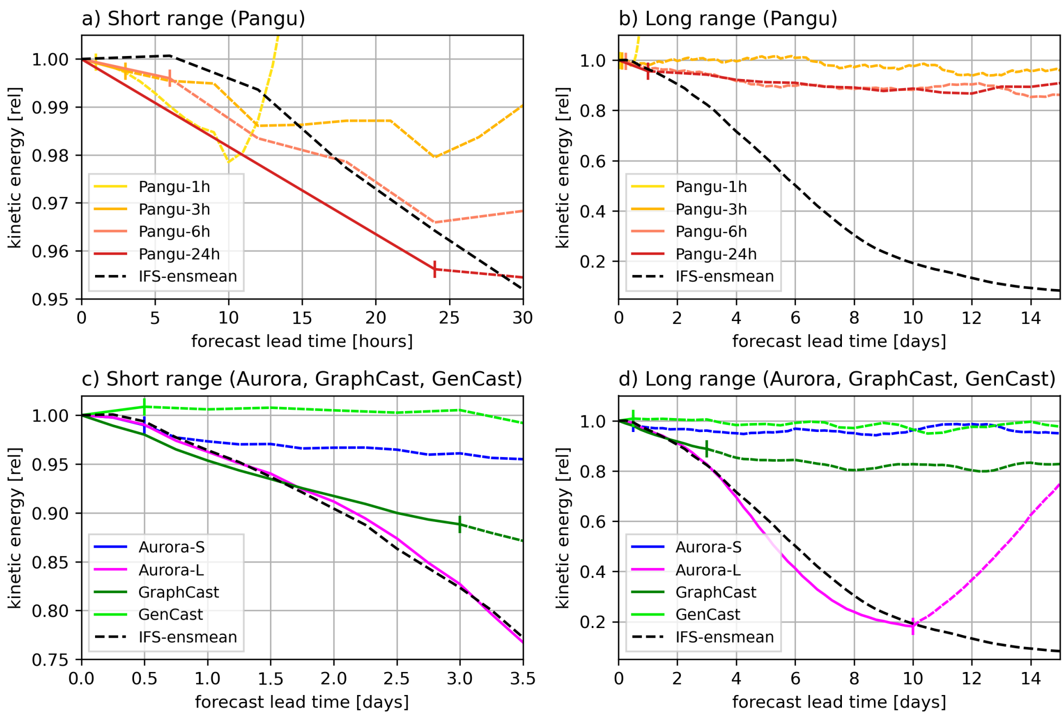

Figure 3. 300 hPa kinetic energy between 400 km and 4000 km wavelength over lead time, relative to initial condition. The plots on the left show a zoom into the initial period. Top and bottom rows show different sets of models. Solid lines indicate lead times within the training interval ($\tau \leq \tau_{\mathrm{train}}$), and dashed lines indicate later lead times. A vertical bar is marking τ_{train}.

but slightly too low. Most notably, the 24-h model at its first time step has a much lower resolution compared to the other three models, which are roughly similar. After the training interval, the 3-h, 6-h, and 24-h model further lose some kinetic energy, but after a few days show a more stable spectrum. The 1-h model however, after an initial loss of kinetic energy, quickly becomes unstable.

For the two Aurora models, Fig. 3 confirms the findings already discussed above: Aurora-S creates a basically stable spectrum, slightly below the IFS-ensemble mean value at the end of the 12-h training interval, while Aurora-L produces a constantly decaying spectrum, closely following the IFS ensemble mean over the 10-day training interval. Note however, that the kinetic energy of Aurora-L is increasing again after this 10-day period, which points to an accumulation of unphysical noise and indicates an unstable model that is not suitable for longer forecasts.

The GraphCast model with its 3-day training interval only roughly follows the IFS ensemble mean, being slightly smoother for the first 1.5 days, and less smooth for the second 1.5 days. This latter behavior contradicts our expectations by producing a forecast with higher effective resolution than the ensemble mean. However, GraphCast was trained using a curriculum approach in which training stated with a single time interval and then slowly increased the lead time interval out to three days. This combined with the fact that GraphCast is a relatively small model is likely lead to the behavior noted above. After the 3 days there is some further decay of kinetic energy, but the spectrum remains stable after about 6-7 days.

GenCast, which is not trained to approximate the ensemble mean or median, but to generate samples from the full distribution, is best able to retain the initial spectrum at all lead times.

3.3 Impact of the resolution on mean-square error scores

A standard way to evaluate the quality of deterministic weather forecasts is to compute the spatially averaged squared difference of some variable to a representation of the truth, referred to as mean-square error. Among others, Ben Bouallègue et al. (2024) demonstrated, that smooth ("low activity") forecasts can lead to better MSE scores by avoiding the double-penalty effect. With the help of the kinetic energy spectrum, we formally explain the reason for the double-penalty effect and confirm it with our simulation data.

An area-weighted mean-square error over the entire globe can equally be computed from spherical harmonics expansions, since Parseval's identity applies. This allows for a scale-dependent formulation of the error, which for error kinetic energy (EKE) reads

$$\mathrm{EKE}(l) = \frac{r^2}{2l(l+1)} \sum_{m=-l}^{l} \Big(|\hat{\zeta}_{lm} - \zeta_{lm}|^2 + |\hat{d}_{lm} - d_{lm}|^2 \Big), \tag{11}$$

where the hat indicates the forecast and non-hat symbols indicate the truth (a similar formalism can be applied to limited domains using Fourier or Cosine transforms). The scale-dependent EKE of the 10-day forecasts is plotted in Fig. 4a, normalized with the kinetic energy (10) of the analysis. For reference, the equally normalized kinetic energy spectrum is shown in Fig. 4b.

To interpret these plots and to understand the double-penalty effect, we expand the absolute square difference,

$$\sum_m |\hat{\zeta}_{lm} - \zeta_{lm}|^2 = \sum_m \Big[(\hat{r}_{lm} - r_{lm})^2 + 2\hat{r}_{lm} r_{lm} \big(1 - \cos(\hat{\phi}_{lm} - \phi_{lm})\big) \Big], \tag{12}$$

where r_{lm} and ϕ_{lm} are amplitude and phase of the complex number ζ_{lm}, respectively. A similar expression holds for any other variable. Consider a mode l, that is no longer

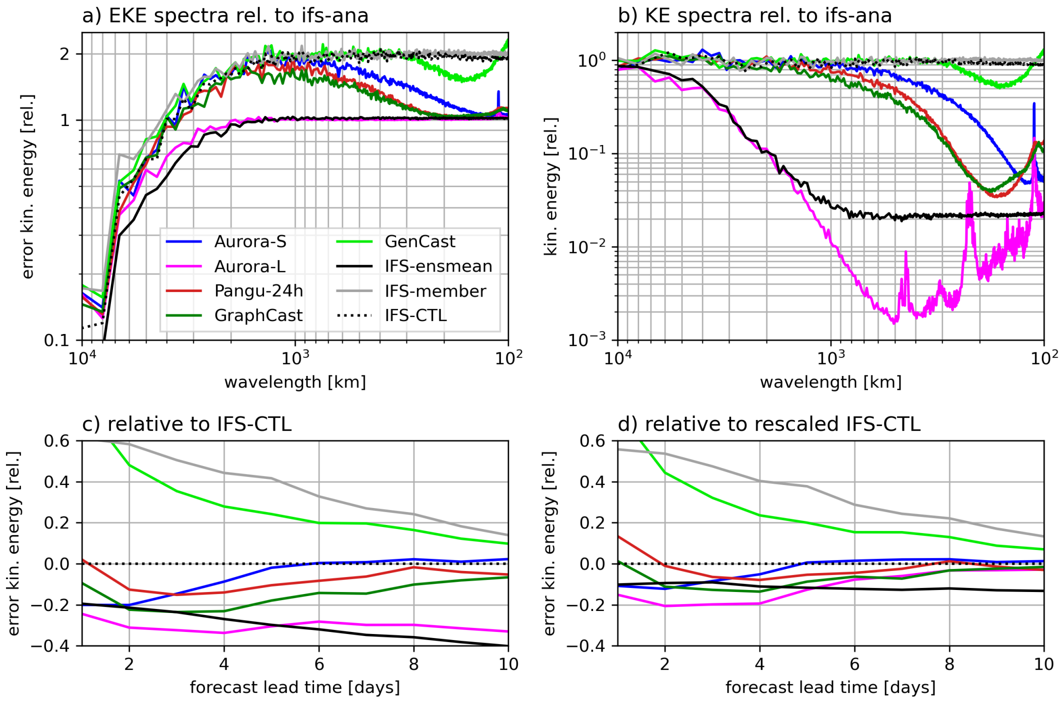

Figure 4. (a) Error kinetic energy spectra of 10-day forecasts over wavelength, relative to the kinetic energy spectrum of the IFS analysis. (b) Same as a, but for kinetic energy spectra. (c) Globally averaged EKE relative to IFS-CTL, computed using (11) and summing over l. (d) Same as c, but relative to a rescaled version of the IFS-CTL by applying (13). Note that these rescale factors differ, depending on the model IFS-CTL was compared to.

predictable. If the model returns zero for that mode, the second term on the left hand side in (12) vanishes and the error equals the amplitude of the analysis spectrum. On the other hand, if the model maintains the correct amplitude but predicts a random phase, the first term vanishes and the error equals *twice* the analysis spectrum (since the expectation of the cosine term is zero) and therefore twice the error compared to predicting zeros (hence double-penalty).

This relation between the error (EKE) and the amplitude (KE) for unpredictable modes becomes evident from our data by comparing Figs. 4a and b: Aurora-L and the IFS ensemble mean produce a normalized EKE of one for scales smaller than 2000 km and at the same time an amplitude close to zero. The other models resemble the IFS-CTL for scales larger than around 1000 km, producing an EKE of two, but an almost realistic amplitude. Towards small scales, the normalized EKE of all AI models except GenCast drops to one as a consequence of their decaying KE. The consequence of the double-penalty effect can also clearly be seen in the EKE time series (Fig. 4c), where smooth forecasts (IFS ensemble mean and Aurora-L) clearly outperform the IFS-CTL and every other model, most significantly at long lead times.

As demonstrated, the scores of the AI models are enhanced by the cancellation of unpredictable modes, which does not indicate a "true" advantage. But the question remains, to what extent? One possibility to exclude the smoothing benefit from a comparison of two models is to equalize their spectra before calculating the EKE or any other mean square error. This can be done by rescaling (damping) the spectral modes of model B to the amplitude of the smoother model A, i.e.,

$$\zeta_{lm}^{\mathrm{B}} \longrightarrow \sqrt{\frac{\sum_m |\zeta_{lm}^{\mathrm{A}}|^2}{\sum_m |\zeta_{lm}^{\mathrm{B}}|^2}}\, \zeta_{lm}^{\mathrm{B}}, \tag{13}$$

and similarly for other variables.

The result of such a comparison is shown in Fig. 4d, where the IFS-CTL spectrum was rescaled to the AI model spectrum. One can see, that the superior skills of the IFS ensemble mean and Aurora-L from Fig. 4c are greatly reduced, especially at long lead times. Indeed for lead times greater than about one week, all AI models perform equally well compared to IFS-CTL, or rather equally badly since there is little practical predictability remaining (Buizza & Leutbecher, 2015; Selz et al., 2022). The difference between Figs. 4c and d is directly correlated to the amount of smoothing produced by the models: It is large for the IFS ensemble mean and Aurora-L, but small for models that approximately maintain the KE spectrum, like Aurora-S, Pangu and GenCast. Note that GenCast is trained to generate samples of the predictive distribution and hence introduces perturbations, which lead to larger errors, especially at early lead times. An even slightly worse degradation of the EKE can be seen from an individual member of the IFS ensemble.

4 Discussion

In summary, we demonstrated with a mathematical argument that the lead time interval in the loss function crucially determines the kinetic energy spectrum and hence the effective resolution of an AI model. If perfectly trained, a model would follow the spectrum of an ideal ensemble mean over that interval and continuously drop unpredictable modes, leading to increasingly smooth forecasts. We also confirmed that smooth forecasts produce much better mean-square error scores by avoiding the double penalty effect and we suggested a method to correct for that.

From our findings, we can identify two basic approaches to weather forecasting with AI: Either a model could be designed to generate samples from the predictive distribution, in which case the lead time training interval should be kept as short as possible. Alternatively, a model could be designed to generate the expectation (the ensemble mean) of the predictive distribution, in which case the lead time training interval should extend to the entire intended forecast lead time.

Both approaches have their justification, however, they should not be mixed and it should be made clear, which one was chosen, since this has consequences for the usage of the model. Models of the first type (Aurora-S, Pangu) can be used to sample the forecast distribution by means of an ensemble, stated from an initial condition sample or using intrinsic stochastisity (GenCast). Each simulation resembles a possible state of the atmosphere that, at least approximately, is physically consistent. Models of the second type on the other hand (like Aurora-L) are not suitable to generate ensembles, do not produce possible realizations of the atmospheric flow and their output is physically inconsistent. However, they do resemble the remaining predictable structures in a single run and predictability can be inferred from the remaining spatial scales.

Although the lead time training interval is crucial for the model's effective resolution, it cannot explain every aspect of it. Most importantly, the presented mathematical argument does not hold for predictions outside of the lead time training interval. In this case, previous forecasts are being fed into the model, which resemble the mean and are much smoother than the training dataset. The reaction of the model to this inconsistency is largely unconstrained. In fact, some models in our study show instability after the lead time training interval and most models continue to lose some kinetic energy. In addition, no model was able to maintain the kinetic energy spectrum on scales smaller than about 300-400 km. Potential reasons for these effects include insufficient training, insufficient capacity, limited sample size of the training data or limitations in the design of the network.

Open Research Section

The AI-model weights, example code and documentation can be found on github: `https://github.com/google-deepmind/GraphCast`, `https://github.com/microsoft/aurora`, `https://github.com/198808xc/Pangu-Weather`. The spherical harmonic coefficients of the forecast data are available at `https://opendata.physik.lmu.de/H66gKyhITQ7qS51` (permanent link after acceptance). The IFS operational analyses, the IFS-CTL and IFS-ENS forecast were retrieved from ECMWF's operational archive (`https://apps.ecmwf.int/archive-catalogue/?class=od`). To obtain access, visit `https://www.ecmwf.int/en/forecasts/accessing-forecasts` for further information.

Acknowledgments

Richard E. Turner is supported by the EPSRC Probabilistic AI Hub and an EPSRC Prosperity Partnership with Microsoft Research. Stratis Markou acknowledges funding from the Vice Chancellor's and George & Marie Vergottis scholarship and the Qualcomm Innovation Fellowship. Anna Vaughan is supported by the UKRI Centre for Doctoral Training in Application of Artificial Intelligence to the study of Environmental Risks (AI4ER). The use of ECMWF's computing and archive facilities is gratefully acknowledged.

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.

Ben Bouallègue, Z., Clare, M. C., Magnusson, L., Gascon, E., Maier-Gerber, M., Janoušek, M., … others (2024). The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an operational-like context. *Bulletin of the American Meteorological Society*, *105*(6), E864–E883.

Ben Bouallègue, Z., et al. (2024). Accuracy versus activity. *AIFS Blog*. doi: 10.21957/8b50609a0f

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.

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Brandstetter, J., Garvan, P., ... others (2024). Aurora: A foundation model of the atmosphere. *arXiv preprint arXiv:2405.13063*.

Bonavita, M. (2024). On some limitations of current machine learning weather prediction models. *Geophysical Research Letters*, *51*(12), e2023GL107377.

Buizza, R., & Leutbecher, M. (2015). The forecast skill horizon. *Quarterly Journal of the Royal Meteorological Society*, *141*(693), 3366–3382.

Charlton-Perez, A. J., Dacre, H. F., Driscoll, S., Gray, S. L., Harvey, B., Harvey, N. J., ... others (2024). Do ai models produce better weather forecasts than physics-based models? a quantitative evaluation case study of storm ciarán. *npj Climate and Atmospheric Science*, *7*(1), 93.

Chattopadhyay, A., Sun, Y. Q., & Hassanzadeh, P. (2024). Challenges of learning multi-scale dynamics with ai weather models: Implications for stability and one solution. Retrieved from `https://arxiv.org/abs/2304.07029`

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... Thépaut, J.-N. (2017). Complete era5 from 1940: Fifth generation of ecmwf atmospheric reanalyses of the global climate [dataset]. *Copernicus Climate Change Service (C3S) Data Store (CDS).* (Accessed on 12-Jul-2023) doi: 10.24381/cds.143582cf

Hsieh, W. W. (2023). *Introduction to environmental data science.* Cambridge University Press.

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., ... others (2024). Neural general circulation models for weather and climate. *Nature*, *632*(8027), 1060–1066.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., ... others (2023). Learning skillful medium-range global weather forecasting. *Science*, *382*(6677), 1416–1421.

Nastrom, G., & Gage, K. (1985). A climatology of atmospheric wavenumber spectra of wind and temperature observed by commercial aircraft. *J. Atmos. Sci*, *42*, 950–960.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., ... others (2025). Probabilistic weather forecasting with machine learning. *Nature*, *637*(8044), 84–90.

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F. A., ... Courville, A. (2019). *On the spectral bias of neural networks.* Retrieved from `https://arxiv.org/abs/1806.08734`

Schulzweida, U. (2024). *Cdo user guide.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.7112925` doi: 10.5281/zenodo.7112925

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

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.

Xu, L., Wang, S., & Tang, R. (2019). Probabilistic load forecasting for buildings considering weather forecasting uncertainty and uncertain peak load. *Applied energy*, *237*, 180–195.