

---

# TESTING THE LIMIT OF ATMOSPHERIC PREDICTABILITY WITH A MACHINE LEARNING WEATHER MODEL

---

A PREPRINT

**P. Trent Vonich** <sup>\*1,2</sup> and **Gregory J. Hakim** <sup>†1</sup>

<sup>1</sup>Department of Atmospheric Sciences, University of Washington, Seattle, WA, USA

<sup>2</sup>Air Force Institute of Technology, Wright-Patterson AFB, OH, USA

April 30, 2025

## ABSTRACT

Atmospheric predictability research has long held that the limit of skillful deterministic weather forecasts is about 14 days. We challenge this limit using GraphCast, a machine-learning weather model, by optimizing forecast initial conditions using gradient-based techniques for twice-daily forecasts spanning 2020. This approach yields an average error reduction of 86% at 10 days, with skill lasting beyond 30 days. Mean optimal initial-condition perturbations reveal large-scale, spatially coherent corrections to ERA5, primarily reflecting an intensification of the Hadley circulation. Forecasts using GraphCast-optimal initial conditions in the Pangu-Weather model achieve a 21% error reduction, peaking at 4 days, indicating that analysis corrections reflect a combination of both model bias and a reduction in analysis error. These results demonstrate that, given accurate initial conditions, skillful deterministic forecasts are consistently achievable far beyond two weeks, challenging long-standing assumptions about the limits of atmospheric predictability.

## 1 Introduction

For more than half a century, atmospheric predictability has been framed by Lorenz’s seminal concept of the “butterfly effect,” which proposes that infinitesimal errors in initial conditions grow rapidly, ultimately limiting skillful deterministic weather forecasts to approximately two weeks [Lorenz, 1969]. This paradigm has profoundly shaped meteorological science, fostering the prevailing view that chaos imposes an insurmountable boundary on weather forecasting in the absence of other sources of skill (e.g., the ocean).

Although frequently linked to Lorenz, the two-week predictability limit actually originates from Charney et al. [1966], who reported a 5-day doubling time of errors in a first-generation general circulation model. Extrapolation of these findings suggested that the intrinsic predictability limit for Earth’s atmosphere is about two weeks. This view has widely influenced scientific and public expectations of weather model performance. However, recently Shen et al. [2022, 2023, 2024] clarify that while the original Lorenz 1969 model effectively illustrates chaotic dynamics, it is ill-suited for quantifying the atmosphere’s intrinsic predictability due to its absence of baroclinic and dissipative processes—a critique Lorenz would later acknowledge himself [Lorenz, 1996].

Modern experiments utilizing models that do include these processes have increased the intrinsic limit modestly beyond two weeks. Zhang et al. [2019] find that reduction of current-day initial condition error by an order of magnitude would yield mid-latitude forecast gains to 15 days. Similarly, “perfect twin” experiments with convection-allowing models show slightly longer limits, with errors plateauing at 17 days in the mid-latitudes and beyond 20 in the tropics [Judt, 2018, 2020]. Selz [2019] reiterate a 17-day limit using the ICON (Icosahedral Nonhydrostatic) model. Finally, while not deterministic, analyses of real-world ECMWF ensembles have offered the most optimistic outlook, demonstrating marginal skill up to 23 days [Buizza and Leutbecher, 2015].

---

<sup>\*</sup>[tvonich@uw.edu](mailto:tvonich@uw.edu)

<sup>†</sup>[ghakim@uw.edu](mailto:ghakim@uw.edu)

The emergence of machine learning (ML) weather models provides a new tool to assess predictability and reduce errors by adjusting forecast initial conditions. As demonstrated in Vonich and Hakim [2024], the backpropagation and gradient descent techniques used to train models may also be used to create an optimal initial condition (visual depiction in Fig. S1), defined as the input that best reproduces a target sequence. This method resembles classical adjoint approaches [e.g., Langland et al., 1995, 2002, Doyle et al., 2012, 2014, 2019, Lloveras et al., 2025] except that it does not require explicit linearization [Vonich and Hakim, 2024, Baño-Medina et al., 2025].

Using the GraphCast model [Lam et al., 2023] and applying this method to the June 2021 Pacific Northwest heatwave [Thompson et al., 2022, Leach et al., 2024], Vonich and Hakim [2024] show that the optimized forecast achieves an 85% reduction in 10-day error compared to a control originating from an ERA5 (ECMWF Reanalysis v5) initial condition [Hersbach et al., 2020], with improvements lasting to 22.5 days. Moreover, forecasts with a different model (Pangu-Weather; [Bi et al., 2023]) initialized with the GraphCast-optimized inputs show comparable 10-day forecast improvements, suggesting that model error is not a critical component of the optimal initial condition. In order to assess the generality of the findings from this case study, we increase the sample size to address three questions:

1. How consistently does initial condition optimization enhance forecast accuracy?
2. What is the maximum lead time for which forecast skill can be achieved with this approach?
3. How reliably can optimized initial conditions produced by GraphCast improve predictions in a different model?

We refine the method described in Vonich and Hakim [2024] and apply it to 732 unique initialization times—forecasts generated at 00Z and 12Z for every day of 2020—and verify the outputs against ERA5. Results show 10-day forecast improvements that are similar in magnitude to that of the 2021 heatwave study (86%) and forecast skill that is about double the current estimates of intrinsic predictability. When tested in Pangu-Weather, these optimized initial conditions still yield statistically significant—though smaller—improvements, suggesting both genuine reduction of initial-condition error and model-specific bias correction.

## 2 Forecast Performance

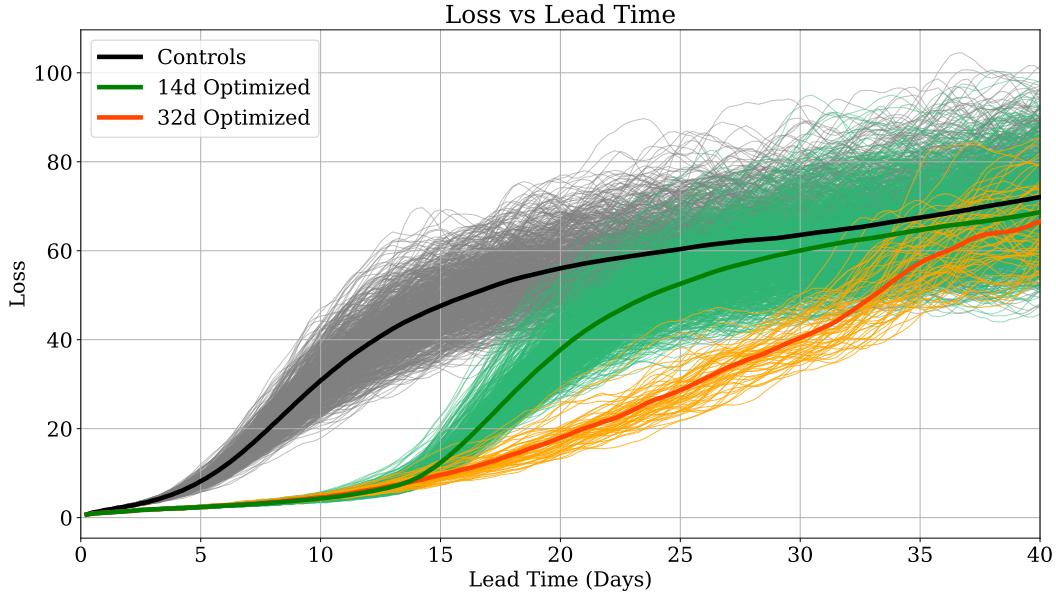


Figure 1: Weighted mean squared error as detailed in Eq. 2 for all 732 control forecasts (black), 14-day optimized forecasts (green), and 32-day optimized forecasts (orange) during 2020.

Each initial condition is optimized per the GraphCast loss function (Sec. 6.3) to reduce cumulative global forecast error over a 14-day window, yielding a set of 732 optimized forecasts computed using 32-bit floating-point arithmetic (hereafter, “single-precision”). We restrict the optimization to 14 days due to diminishing returns stemming from loss of numerical precision, as longer windows require increasingly fine adjustments to the initial condition. To explore

extended forecast horizons, we compute a 61-member subset—every sixth day of 2020—using 64-bit floating-point arithmetic (hereafter, “double-precision”). GPU memory constraints limit double-precision optimization to 32 days, but there is no indication that the process could not continue further with sufficient computing resource.

When the loss is measured at ten days, the single-precision mean (green) shown in Fig. 1 displays an 86% reduction in error compared to the control mean (black). Surprisingly, there are no failures. Each initialization time can be substantially optimized, with a minimum improvement of 77% and a maximum of 91%. Given that all 732 forecasts exhibit considerable error reduction up to 14 days, the technique appears effective for a wide range of atmospheric states. Beyond the 14th day, rapid error growth resumes. The optimization does not assimilate information beyond this time, so error growth returns to a rate that mirrors the control at earlier times until the two eventually merge near 30 days.

The double-precision results (orange), optimized to 32 days, show a reduction in error relative to the control forecasts after the single-precision optimizations fail. The control and double-precision sample means have approximately equal error at 5 and 15 days, respectively. Errors grow at a nearly uniform exponential rate, with a doubling time of 5.8 days from day 2.5 until day 14 for both the single and double-precision sample means. Ultimately, the double-precision error growth rate gradually decreases as it merges with the control around 37 days. Like the single-precision results, the double-precision curve also exhibits a subtle increase in error growth rate after the assimilation window ends (day 32). The constant doubling time of errors is clearer when plotted with a logarithmic y-axis (Figure S2), which also reveals an initial phase of elevated error growth (average doubling time of  $\sim 1 \text{ day}^{-1}$ ) between 6 and 24 hours, followed by a deceleration and a transition to the persistent 5.8-day doubling rate.

It is worth noting that for exceptionally long forecasts—beyond 45 days—GraphCast is known to become unstable [Karl Bauer et al., 2024], and the results show early evidence of this in Fig. 1. The mean loss curves show a modest upward slope beyond 35+ days, never fully saturating. As a result, it is not clear from the loss exactly where forecast skill is lost, so we compute the anomaly correlation coefficient (ACC) using the WeatherBenchX library [Rasp et al., 2023] and find that for Z500 the anomaly correlation remains statistically different from the control at  $p \leq 0.01$  to 33 days (Fig. S3). Practical forecast skill, commonly defined as an ACC of 0.6 by ECMWF [Zhang et al., 2019], persists to 27.5 days.

### 3 Optimal Perturbation Sample-Mean Structure

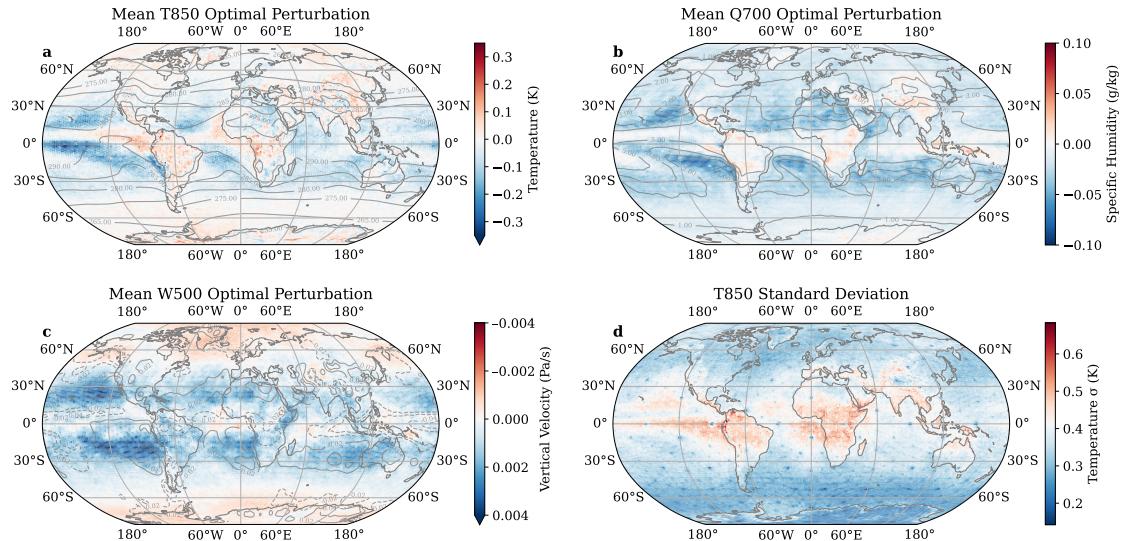


Figure 2: Sample-mean optimal perturbations averaged over 732 cases for (a) 850 hPa temperature, (b) 700 hPa specific humidity, (c) 500 hPa pressure vertical velocity (negative values indicate rising air); and (d) 850 hPa temperature sample standard deviation. Gray solid (dashed) contours represent the corresponding positive (negative) sample-mean values for ERA5.

The sample-mean optimal perturbations reveal coherent large-scale structure with greatest amplitude in the tropics and subtropics (Fig. 2). At 850 hPa, temperature perturbations (Fig. 2a) exhibit hemispherically symmetric cold anomalies over regions of subtropical stratocumulus cloud decks with warm anomalies along the Intertropical Convergence Zone

(ITCZ), off the coast of Ecuador, and along the African west coast. Cooling along the equatorial east Pacific may also be related to the 2020 La Niña event, possibly capturing the eastward extension of this event’s cold tongue [Li et al., 2022].

The 700 hPa specific humidity perturbations (Fig. 2b) correspond spatially to the temperature perturbations, showing subsidence-driven drying of the mid-troposphere across subtropical oceans and moistening near the ITCZ, Central America, sub-Saharan Africa, and China. The central Indian Ocean and the Maritime Continent also exhibit increased moisture, consistent with the westward-shifted warm pool during La Niña [Li et al., 2022]. Pressure vertical velocity perturbations at 500 hPa (Fig. 2c) further highlight the coherent structure of the GraphCast optimal initial conditions, with enhanced upward motion near the ITCZ, and increased subsidence throughout the subtropics. Increased upward motion relative to ERA5 characterizes the polar regions.

To put the perturbation amplitude range in perspective, Fig. 2d shows the 850 hPa temperature standard deviation. The most active regions generally mirror Fig. 2a, showing that perturbations along the ITCZ, northern South America, central Africa, and the Maritime continent have the greatest mean magnitude relative to ERA5. This may suggest that these regions are more poorly resolved than others or could reflect regional biases within GraphCast. Supplementary perturbation statistics reveal that the average magnitude of the perturbations is on the order of typical analysis error for all variables [e.g., Daley and Mayer, 1986, Hakim, 2005, Peña and Toth, 2014] (Table 9). Additionally, the icosahedral vertices and edges of the graph neural network linger as visible artifacts within the standard deviation data. They induce a locally smaller standard deviation and are likely tied to GraphCast’s encoding and decoding layers [Lam et al., 2023].

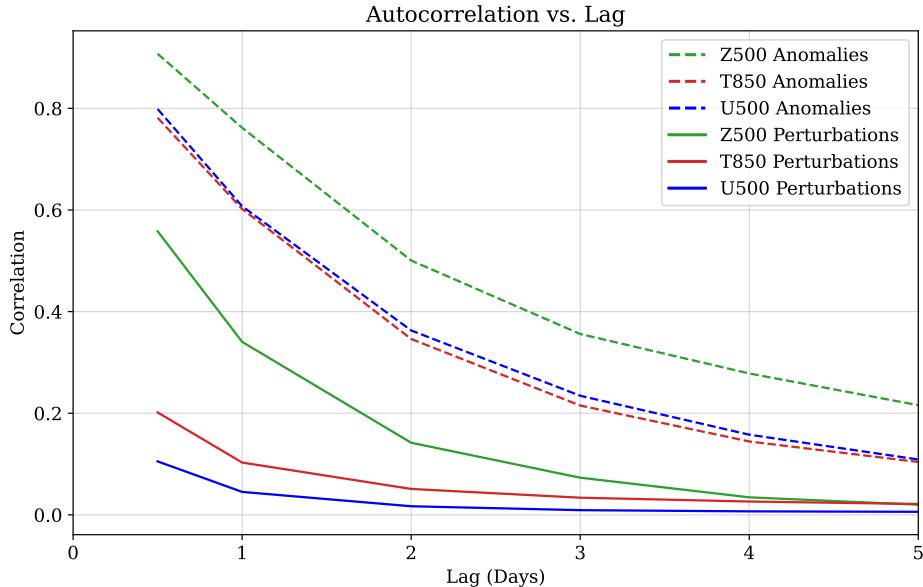


Figure 3: ERA5 2020 climatological anomaly and optimal perturbation autocorrelation as a function of lag for geopotential (Z500), temperature (T850), and zonal wind (U500) at their respective pressure levels. Solid lines represent the global mean autocorrelation for the optimal perturbations while dashed lines show the ERA5 anomaly global mean autocorrelation. The first correlation value is computed at 12 hours, consistent with the twice-daily optimization.

Overall, the sample-mean optimal structure represents a strengthening of the Hadley circulation, consistent with the weaker divergent wind component in ERA5 [Li et al., 2024]. Analysis of the sample-mean perturbations in time and space reveals a distinct autocorrelation for each variable. In the global average, geopotential height has the most persistent autocorrelation for the optimal perturbations (see Fig. 3). Temperature and zonal wind exhibit substantially lower initial autocorrelation values compared to geopotential, but all three display similar e-folding times of 1.0 to 1.5 days. Zonal wind and specific humidity (not shown) show enhanced autocorrelation in the tropics, whereas geopotential height has a more spatially uniform pattern. These results suggest that certain components of the optimized initial conditions—particularly the geopotential height—may be exploitable for operational forecasting. However, the rapid decay in autocorrelation, especially for temperature and wind, indicates that a substantial portion of the adjustments are tuned specifically to the atmospheric state at the time of initialization. The ERA5 anomaly mean autocorrelation is included in Fig. 3 for reference and shows that decay rates of the observed anomalies closely parallel those of the perturbations. We also find that simply adding the sample-mean optimal perturbations to the control (ERA5) initial

conditions used in Section 2 reduces the loss by an average of 1-2% over a 30-day period relative to the control forecasts (Fig. S4).

## 4 Cross-Model Forecast Validation

To assess the degree to which the optimal initial conditions encode forecast error from the GraphCast model, we feed all 732 optimized initial conditions into the Pangu-Weather model [Bi et al., 2023]. Pangu-Weather is chosen for its distinctly different architecture, inference method, and spatial resolution. Unlike GraphCast, which uses two 6-hour time steps for autoregressive forecasts, Pangu-Weather predicts 24-hour steps from a single time input, does not have vertical velocity and precipitation inputs, and operates at higher spatial resolution ( $0.25^\circ$ ), necessitating spherical harmonic interpolation of the optimized inputs.

Fig. 4 reveals a marked improvement for 500 hPa geopotential height throughout the 14-day optimization window, but greater variability (some forecasts are worse than the control) and generally smaller improvement relative to GraphCast. When measured at 4 days, GraphCast optimal forecasts average a 62% loss reduction whereas Pangu-Weather forecasts only average 21%, crudely suggesting a 2:1 ratio of model error to initial condition error. Nevertheless, the best Pangu-Weather forecasts show a 60% loss reduction at 10 days and 50% reduction to 30 days—comparable in magnitude to the mean GraphCast results.

Several factors likely contribute to smaller forecast improvements with Pangu-Weather. First, model-specific biases inherent to Pangu-Weather are likely different from GraphCast, diluting the impact of the optimized inputs. Second, GraphCast’s two 6-hour time-steps allow refinement of short-term tendencies, whereas Pangu-Weather—designed for single-step 24-hour forecasts—can only accept one of these optimized time levels, effectively receiving only half the information contained in the full optimization. Third, the absence of vertical velocity and precipitation in Pangu-Weather’s input also reduces the available optimized information. Finally, the interpolation process needed to project the optimized state onto the  $0.25^\circ$  grid introduces errors that may degrade the forecast performance given Pangu-Weather’s finer resolution. Despite these drawbacks, we note that a 21% error reduction represents roughly a decade of forecast skill advancement at the current pace of operational improvement [Bauer, 2024].

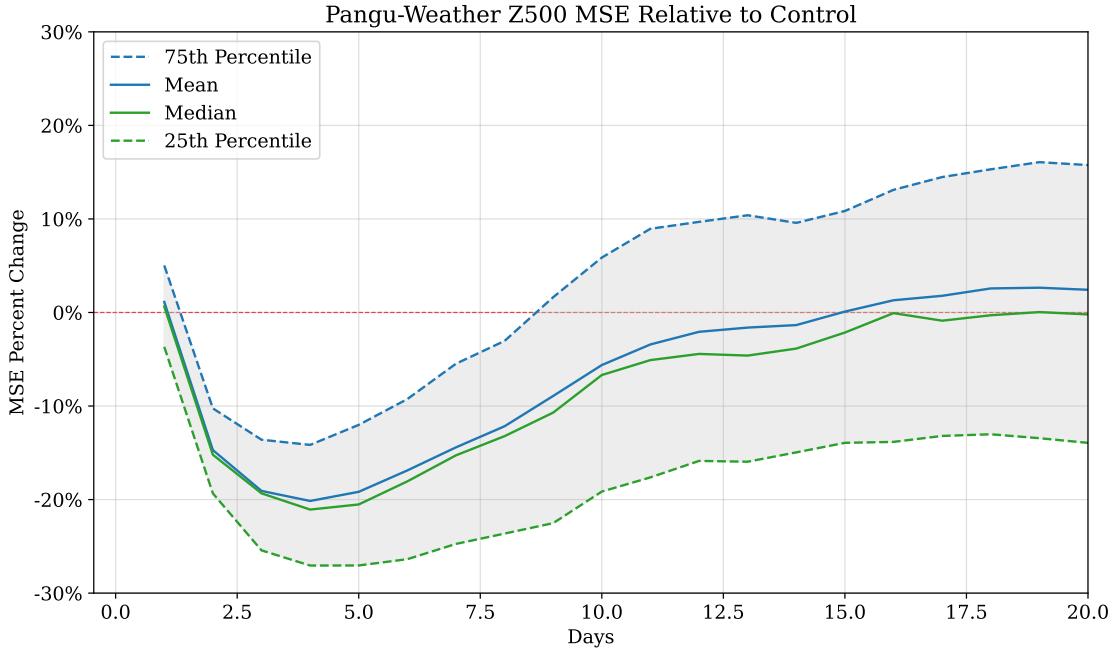


Figure 4: Mean, median, and interquartile range (25th–75th percentiles) of the relative change in mean squared error for Pangu-Weather forecasts using 732 GraphCast-optimized initial conditions. The control forecasts use  $1.0^\circ$  ERA5 data interpolated to a  $0.25^\circ$  grid for an equal comparison.

## 5 Discussion and Conclusion

Using gradient-based optimization of initial conditions with the GraphCast model, we find weather forecast skill lasting twice as long as the hitherto established limit of atmospheric predictability, exhibiting statistical significance to 33 days and useful skill up to 27.5 days. Cross-model validation with the Pangu-Weather model confirms that the optimized initial conditions yield significant, albeit smaller, improvements, suggesting that the GraphCast-optimized initial conditions involve a blend of analysis improvement and model-specific bias correction. Since GraphCast was trained on ERA5, separating model bias from reanalysis error remains ambiguous. The sample-mean optimal perturbations exhibit spatially coherent adjustments to the ERA5 reanalysis that broadly reflect an intensification of the Hadley circulation. We emphasize that traditional predictability studies have typically ascribed an intrinsic predictability limit by identifying the time at which two arbitrarily similar initial states become climatologically indistinguishable. In contrast, our study defines the limit for a single deterministic forecast as the time beyond which adjustments to the initial condition no longer reduce the error when verified against ERA5.

A potential concern with the optimization procedure is that it may represent a form of adversarial attack on the neural network [e.g., Szegedy et al., 2014, Moosavi-Dezfooli et al., 2016, Tabacof and Valle, 2015]. These attacks have been observed to change the output categorization of image classifier models through subtle single-pixel attacks [Goodfellow et al., 2015]. Adversarial examples are known to transfer across models under certain circumstances, meaning that testing results on different architectures (e.g., Pangu-Weather) may not always be protective. However, recent research has reframed these vulnerabilities, suggesting that what were once considered bugs are actually features of the model [Ilyas et al., 2019, Springer et al., 2021]. Although brittle, the pathways on which adversarial attacks capitalize, remain predictive. In any case, we hypothesize that several aspects of our method make it less prone to such attacks.

First, in contrast to adversarial attacks, which aim to maximize output loss, our procedure seeks to minimize it. Second, across 732 distinct forecasts, a clear lower loss bound emerges: continued optimization—even using 10 times as many epochs—does not yield appreciable improvement. Moreover, we note that when initializing the optimization of one ERA5 trajectory (1–14 January 2020) with the wrong starting point (a random sample from climatology), the optimization still closely reproduces the optimal ERA5 reanalysis for that date. This suggests that the gradient descent calculation is robust to large errors in the starting point because the loss is minimized over a long trajectory through phase space (i.e., 56 consecutive states), which necessitates a particularly precise initial condition.

Although GraphCast is deterministic—producing a single output for a given input—its predictions become blurred under multi-day loss minimization [Brenowitz et al., 2024, Charlton-Perez et al., 2024, Bonavita, 2024]. As a result, its outputs are not directly comparable to those of traditional deterministic physics-based models. Future work could address this issue by optimizing initial conditions for models engineered to minimize blurring, higher resolution models, and coupled atmosphere–ocean ML models [e.g., Cresswell-Clay et al., 2025]. It would also be valuable to evaluate the error-reducing potential of optimal perturbations designed for physics models, perhaps through nudging as in Husain et al. [2024]. Our findings support the prospect of a longer intrinsic predictability horizon, offering a hopeful outlook for untapped advancement in weather forecasting. By refining initial conditions and probing model behavior, this approach provides a foundation for sustained progress and more skillful prediction of the atmosphere.

## 6 Methods

The methodology employed in this study follows that described in Vonich and Hakim [2024], with a key modification represented in Step 5 of the enumerated optimization process.

### 6.1 Model

We use the “small” version of the GraphCast model [Lam et al., 2023], selected for its modest memory footprint. This enables gradient computations over extended windows, up to 32 days, in a practical timeframe. GraphCast forecasts six atmospheric state variables: geopotential height, temperature, specific humidity, vertical velocity, and zonal and meridional wind components, resolved across 13 pressure levels. It also predicts four surface variables—mean sea-level pressure, 2-meter air temperature, and 10-meter zonal and meridional wind components—alongside 6-hour accumulated precipitation, all on a  $1.0^\circ \times 1.0^\circ$  grid. With 36.7 million parameters, GraphCast was trained on ERA5 reanalysis data from 1979 to 2015 [Hersbach et al., 2017].

To generate predictions, inference utilizes two atmospheric input states, separated by a 6-hour interval, producing a single output state 6 hours in the future. For extended forecasts, the output is autoregressively fed back into the model alongside the prior 6-hour state, enabling indefinite prediction. In this study, optimal perturbations were computed

exclusively for the state variables, while static fields, including the land-sea mask and surface geopotential, remained unaltered.

## 6.2 Optimization

Our approach leverages the fully differentiable nature of GraphCast to optimize initial conditions in a nonlinear framework, overcoming the linear limitations of traditional adjoint models. Unlike adjoint models, which cannot reliably represent atmospheric processes beyond five days due to linearization limits [Langland et al., 1995, Errico, 1997], deep learning models provide a versatile alternative. They integrate linear and nonlinear operations across layers, enabling seamless derivative computation via the chain rule. In this study, all automatic differentiation is performed using GraphCast implemented in the JAX framework [Lam et al., 2023]. JAX provides robust support for automatic differentiation, complemented by GPU acceleration and dynamic code optimization [Bradbury et al., 2018].

This differentiable framework enables iterative refinement of atmospheric initial conditions. Given the input state  $\mathbf{x}_i$  for iteration  $i$  (where  $i = 0$  is the unaltered ERA5 initial condition), we compute an increment based on the gradient of the forecast loss,  $\mathcal{L}(\mathbf{N}(\mathbf{x}_i))$ , with respect to the inputs. Here,  $\mathbf{N}$  denotes GraphCast inference starting from the chosen forecast optimization time.

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{x}_i}. \quad (1)$$

The derivative in (1) entails tracing the loss gradient back through the GraphCast neural network for every 6-hour time step. This gradient characterizes how changes to initial inputs will influence the final forecast error. Adjoint models can also be employed to perform gradient descent, but, like deep learning models, they may struggle to navigate complex gradient landscapes with numerous saddle points and valleys [Pires et al., 1996]. In Figure 4 of Vonich and Hakim [2024], this phenomenon appears in forecast optimizations beyond 5 days due to increasing gradient complexity with longer lead times. To overcome this problem, we gradually expand the optimization window size rather than fitting the entire trajectory at once. This proves to be a simpler gradient descent task and allows the algorithm to smoothly navigate what might be a complex loss manifold. The size of the optimization windows is arbitrary, but we choose an initial length of 2 days to allow forecast error to develop that is distinct from analysis error. Subsequent steps expand the window in 3-day increments, which works effectively based on empirical testing. Swanson et al. [1998] implement a similar strategy for a four-dimensional variational data assimilation solver, also noting the performance improvement offered by progressively assimilating the total available data. They refer to this method as quasi-static, reflecting the stepwise adjustment of the assimilation window.

Details of the algorithm in the production of one set of optimized initial conditions (an “optimal”) proceeds as follows:

1. Given a set of inputs, produce a forecast for the optimization window size. On the first pass, the initial window size is 2 days, and we initialize the forecast with the ERA5 analysis. Every subsequent epoch and window size starts with the optimal computed on the previous step.
2. Calculate the forecast loss function by verifying against ERA5 at every step during inference.
3. Calculate the gradient of the loss function with respect to the inputs using the JAX framework.
4. Update the inputs using the Adam optimizer for gradient descent [Kingma and Ba, 2017], applying the loss gradient as per Equation (1).
  - (a) Repeat steps 1 – 4 for a specified number of epochs, then proceed to step 5.
5. Increase the optimization window size in step 1 by 3 days.
  - (a) Repeat steps 1 – 5 until the maximum optimization window size is reached.

In this study, the maximum optimization window size is 32 days, limited by the 80 GB memory of the NVIDIA A100 GPU. As the forecast duration increases, so does the gradient size, necessitating a smaller model, greater GPU memory, or a strategy to further extend the window size. We suspect that optimization past 32 days would yield modest additional improvement. With respect to the optimizer hyperparameters, the default values ( $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ) and a learning rate of  $10^{-3}$  are used, which have been shown to be effective in our previous work. We progressively reduce the learning rate to handle the increasingly complex gradient descent for optimization windows longer than 14 days. For variables constrained to non-negative values, such as specific humidity and precipitation, optimization occasionally yields small negative perturbations; these are clipped to zero with negligible impact on the results. All findings reported in this paper represent the clipped initial conditions. With end-to-end double-precision floating-point operations, each optimization requires around 4 hours on an NVIDIA A100 GPU.

### 6.3 Loss Function

As in our original work, we adopt the scalar loss function used to train GraphCast, a weighted mean squared error that quantifies the difference between predicted and target outputs, averaged over time, variables, and spatial locations. For a predicted state  $\hat{x}$  and verification state  $x$ , the loss is expressed as:

$$\mathcal{L}_{\text{MSE}} = \underbrace{\frac{1}{T_{\text{time}}} \sum_{\tau=1}^{T_{\text{time}}}_{\text{lead time}}}_{\text{spatial location}} \underbrace{\frac{1}{|G_{1.0^\circ}|} \sum_{i \in G_{1.0^\circ}}}_{\text{variable-level}} \sum_{j \in J} s_j w_j a_i (\hat{x}_{i,j}^{t_0+\tau} - x_{i,j}^{t_0+\tau})^2 \quad (2)$$

In this equation,  $w$  represents the weight by pressure level,  $a$  is the grid-cell area, and  $s$  is a standardization parameter computed from time differences in the GraphCast training data. For more details on these parameters and the loss function, refer to Section 4.2 of the Materials and Methods in Lam et al. [2023].

### 6.4 Statistical Significance

Forecast errors may be temporally correlated, which motivates an estimate of the effective sample size for assessing statistical significance of long-lead forecasts. Therefore, we quantify the autocorrelation of Z500 forecast errors at 35-days, adjust the sample size accordingly using an effective sample size estimate, and compute corrected critical values for statistically significant anomaly correlation coefficients (ACC).

Let  $\{e_{n,L}\}_{n=1}^N$  be the sample of Z500 forecast errors at a fixed lead time  $L$ , where  $N = 61$  is the number of double-precision forecasts and  $L = 35$  days. Define the lag- $k$  autocorrelation at lead  $L$  by

$$r_k(L) = \text{corr}(e_{n,L}, e_{n-k,L}).$$

In particular, for a 6-day lag ( $k = 1$ ) when  $L = 35$  days we observe

$$r_1(35) = 0.08.$$

Moreover, for lead times  $L = 1, 2, \dots, 15$  days, the lag-1 autocorrelation remains below 0.01, indicating that forecast errors at these shorter leads are effectively independent.

To correct for this temporal correlation when testing anomaly correlation coefficients (ACC), we compute the effective sample size [Wilks, 2011]:

$$N_{\text{eff}} = \frac{N(1 - r_1)}{1 + r_1} \approx \frac{61(1 - 0.08)}{1 + 0.08} \approx 52.$$

Using a one-tailed  $t$ -test with  $\nu = N_{\text{eff}} - 1 = 51$  degrees of freedom, the critical ACC values  $r_c$  satisfy

$$r_c = \frac{t_{\alpha,\nu}}{\sqrt{t_{\alpha,\nu}^2 + \nu}},$$

which yields

$$r_c(p \leq 0.05) \approx 0.23, \quad r_c(p \leq 0.01) \approx 0.32.$$

Since our sample has no optimization failures, increasing  $N$  would likely raise  $N_{\text{eff}}$ , thereby marginally extending the maximum lead time for which ACC values are statistically significant.

## 7 Data availability

All ERA5 data used in this study is openly available from the Copernicus Data Store (<https://doi.org/10.24381/cds.143582cf>) Hersbach et al. [2017]. All code required to operate GraphCast and Pangu-Weather can be found at <https://github.com/google-deepmind/graphcast> Lam et al. [2023] and <https://github.com/198808xc/Pangu-Weather> Bi et al. [2023], respectively. All optimal initial conditions will be made available on a public hosting platform at the time of publication.

## 8 Author contribution

PTV and GJH conceived the study. PTV performed all optimization calculations and designed the optimization method. GJH designed the Pangu-Weather experiments, including the spherical harmonics interpolation, and PTV performed the computations. PTV prepared all figures, and wrote the initial draft of the manuscript. PTV and GJH revised the draft to the final version. GJH provided overall scientific guidance.

## 9 Acknowledgments

We acknowledge high-performance computing support from the Casper cluster ([doi.org/10.5065/qx9a-pg09](https://doi.org/10.5065/qx9a-pg09)) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation. The Copernicus Climate Data Store provided access to ERA5. We thank Chris Snyder (NCAR), Matthew Chantry (ECMWF), Christian Lessig (ECMWF), Peter Dueben (ECMWF), and Dominik Stiller (UW) for conversations related to this work. PTV was funded by the Dr. Heather Wilson STEM Fellowship. GJH acknowledges support from NSF awards 2202526 and 2402475, and Heising-Simons Foundation award 2023-4715.

## Supplement

This section contains Supplementary figures and data.

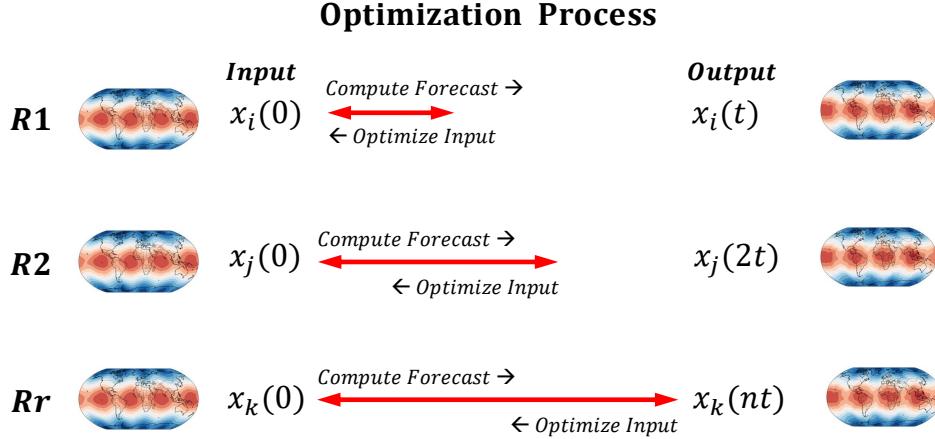


Figure S1: Visualization of the optimization process. A forecast is computed from an initial condition  $x(0)$  for a selected lead time ( $t$ ). The derivative of the global loss function (Eq. 2) is taken with respect to the difference between the computed forecast and a verification dataset (i.e., ERA5). The optimization is repeated  $i$  times and then the process proceeds to Round 2 ( $R2$ ) where the forecast is lengthened by a multiple of  $t$ . See Methods for additional details.

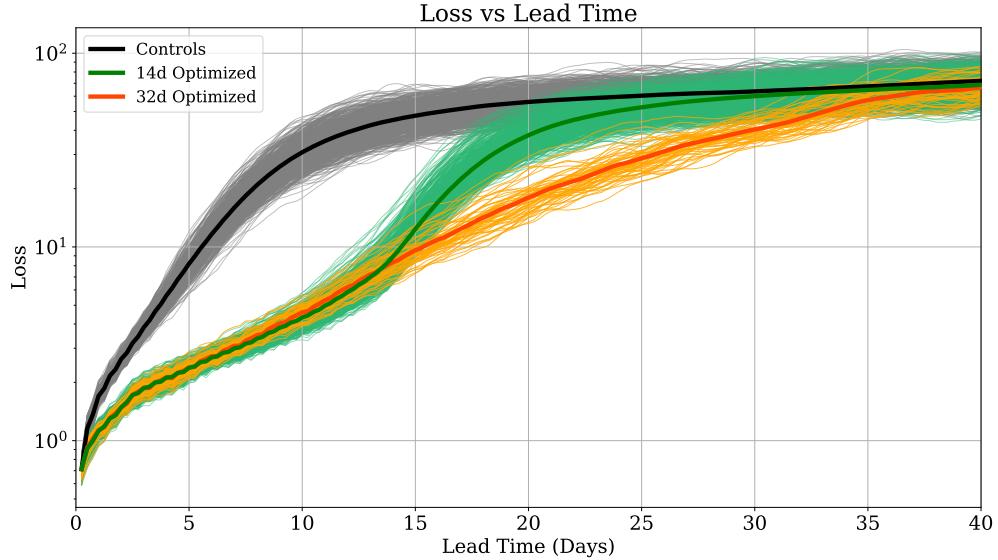


Figure S2: As in Figure 1, but with a log y-axis. Note the initially fast but decelerating error growth between 0 and 2.5 days followed by steady error growth rate thereafter. Recall that the green curves have only been optimized to 14-days and thus return to the control error growth thereafter.

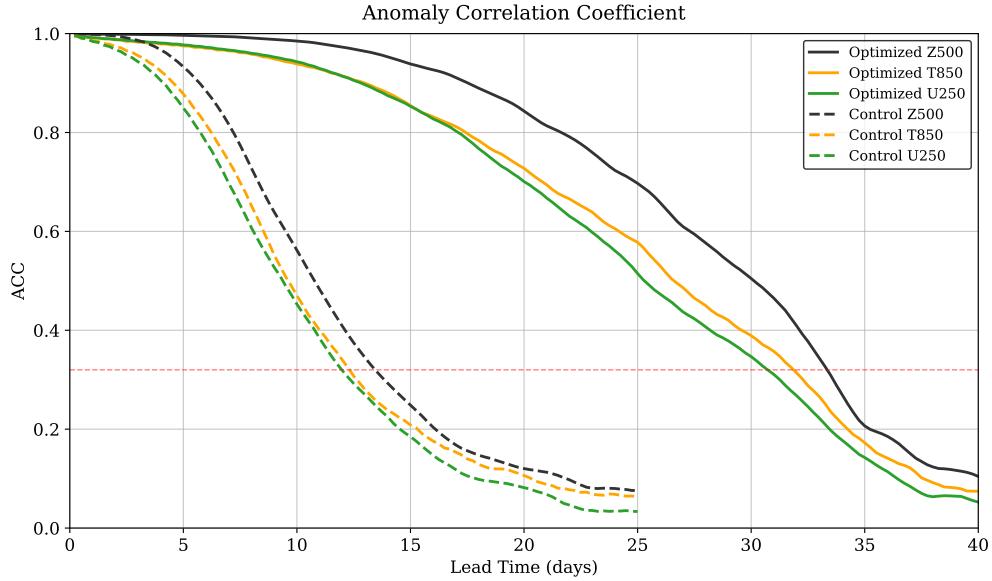


Figure S3: 2020 global-mean anomaly correlation coefficient (ACC) for key variables and pressure levels. Solid lines represent optimal forecasts for geopotential (black), temperature (orange), and zonal wind (green). Dashed lines show corresponding control forecasts for the same variables. The red horizontal dashed line marks the threshold (0.32) at which ACC remains statistically significant ( $p \leq 0.01$ ).

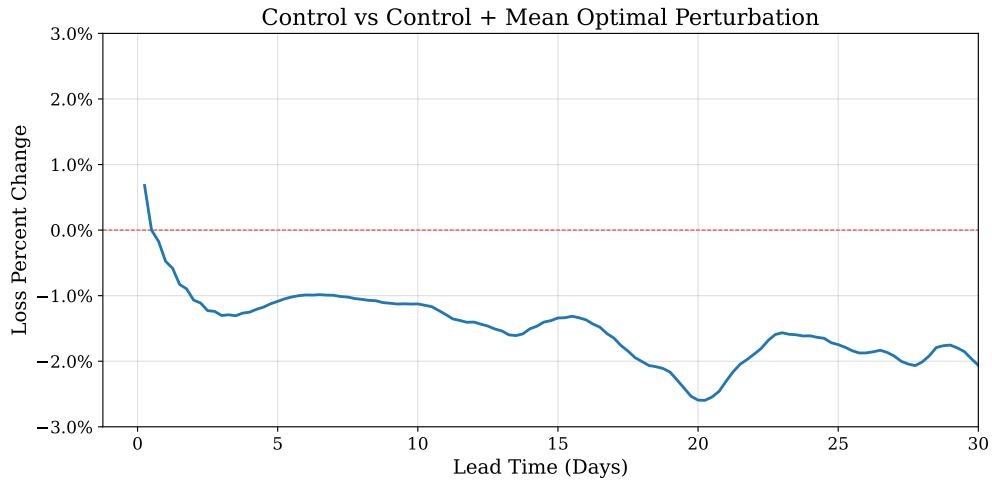


Figure S4: Results for forecasting experiments where the sample-mean optimal perturbation (as illustrated in Fig. 2) is added to the 732 unperturbed initial conditions from ERA5. The forecast loss is then computed, averaged, and compared to the original control loss seen in Fig. 1. The result is a 1–2% average improvement across the 30-day forecast window.

<b>Variable</b>	<b>Mean Magnitude</b>	<b>Perturbation</b>	<b>Mean Standard Devia- tion</b>	<b>Grid Maximum Stan- dard Deviation</b>
200 hPa Zonal Wind	$0.04 \text{ m s}^{-1}$		$0.39 \text{ m s}^{-1}$	$0.88 \text{ m s}^{-1}$
200 hPa Meridional Wind	$0.03 \text{ m s}^{-1}$		$0.28 \text{ m s}^{-1}$	$0.62 \text{ m s}^{-1}$
500 hPa Geopotential Height	$0.63 \text{ m}$		$5.0 \text{ m}$	$10.6 \text{ m}$
500 hPa Pressure Vertical Veloc- ity	$8 \times 10^{-4} \text{ Pa s}^{-1}$		$5 \times 10^{-3} \text{ Pa s}^{-1}$	$1.1 \times 10^{-2} \text{ Pa s}^{-1}$
700 hPa Specific Humidity	$0.02 \text{ g kg}^{-1}$		$0.08 \text{ g kg}^{-1}$	$0.16 \text{ g kg}^{-1}$
850 hPa Temperature	$0.04 \text{ K}$		$0.33 \text{ K}$	$0.68 \text{ K}$

Table 1: Mean absolute value, mean standard deviation, and grid maximum standard deviation for select GraphCast upper-air optimal perturbations. These values are derived from the single-precision optimizations.

## References

- Edward N. Lorenz. The predictability of a flow which possesses many scales of motion. *Tellus*, 21(3):289–307, 1969. doi: 10.1111/j.2153-3490.1969.tb00444.x. URL <http://eaps4.mit.edu/research/Lorenz/publications.htm>.
- J. G. Charney, R. G. Fleagle, V. E. Lally, H. Riehl, and D. Q. Wark. The feasibility of a global observation and analysis experiment. *Bulletin of the American Meteorological Society*, 47:200–220, 1966.
- Bo-Wen Shen, Roger Pielke, Xubin Zeng, Jialin Cui, Sara Faghhi-Naini, Wei Paxson, Amit Kesarkar, Xiping Zeng, and Robert Atlas. The dual nature of chaos and order in the atmosphere. *Atmosphere*, 13(11):1892, 2022. ISSN 2073-4433. doi: 10.3390/atmos13111892. URL <https://www.mdpi.com/2073-4433/13/11/1892>.
- Bo-Wen Shen, Roger A. Pielke, Xubin Zeng, and Xiping Zeng. Lorenz’s view on the predictability limit of the atmosphere. *Encyclopedia*, 3(3):887–899, 2023. ISSN 2673-8392. doi: 10.3390/encyclopedia3030063. URL <https://www.mdpi.com/2673-8392/3/3/63>.
- Bo-Wen Shen, Roger A. Pielke, Sr., Xubin Zeng, and Xiping Zeng. Exploring the origin of the two-week predictability limit: A revisit of lorenz’s predictability studies in the 1960s. *Atmosphere*, 15(7):837, 2024. doi: 10.3390/atmos15070837. URL <https://doi.org/10.3390/atmos15070837>.
- Edward N Lorenz. Predictability: A problem partly solved. In *Proc. Seminar on predictability*, volume 1. Reading, 1996.
- Fuqing Zhang, Y Qiang Sun, Linus Magnusson, Roberto Buizza, Shian-Jiann Lin, Jan-Huey Chen, and Kerry Emanuel. What is the predictability limit of midlatitude weather? *Journal of the Atmospheric Sciences*, 76(4):1077–1091, 2019.
- Falko Judt. Insights into atmospheric predictability through global convection-permitting model simulations. *Journal of the Atmospheric Sciences*, 75(5):1477 – 1497, 2018. doi: 10.1175/JAS-D-17-0343.1. URL <https://journals.ametsoc.org/view/journals/atsc/75/5/jas-d-17-0343.1.xml>.
- Falko Judt. 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, 2020. doi: 10.1175/JAS-D-19-0116.1. URL <https://journals.ametsoc.org/view/journals/atsc/77/1/jas-d-19-0116.1.xml>.
- Tobias Selz. Estimating the intrinsic limit of predictability using a stochastic convection scheme. *Journal of the Atmospheric Sciences*, 76(3):757 – 765, 2019. doi: 10.1175/JAS-D-17-0373.1. URL <https://journals.ametsoc.org/view/journals/atsc/76/3/jas-d-17-0373.1.xml>.
- Roberto Buizza and Martin Leutbecher. The forecast skill horizon. *Quarterly Journal of the Royal Meteorological Society*, 141(693):3366–3382, 2015.
- Peter Vonich and Gregory Hakim. Predictability Limit of the 2021 Pacific Northwest Heatwave from Deep-Learning Sensitivity Analysis, September 2024. URL <https://doi.org/10.5281/zenodo.13694959>.
- Rolf H Langland, Russell L Elsberry, and Ronald M Errico. Evaluation of physical processes in an idealized extratropical cyclone using adjoint sensitivity. *Quarterly Journal of the Royal Meteorological Society*, 121(526):1349–1386, 1995.
- Rolf H. Langland, Melvyn A. Shapiro, and Ronald Gelaro. Initial condition sensitivity and error growth in forecasts of the 25 january 2000 east coast snowstorm. *Monthly Weather Review*, 130(4):957 – 974, 2002. doi: 10.1175/1520-0493(2002)130<0957:ICSAEG>2.0.CO;2. URL [https://journals.ametsoc.org/view/journals/mwre/130/4/1520-0493\\_2002\\_130\\_0957\\_icsaeg\\_2.0.co\\_2.xml](https://journals.ametsoc.org/view/journals/mwre/130/4/1520-0493_2002_130_0957_icsaeg_2.0.co_2.xml).
- James D. Doyle, Carolyn A. Reynolds, Clark Amerault, and Jonathan Moskaitis. Adjoint sensitivity and predictability of tropical cyclogenesis. *Journal of the Atmospheric Sciences*, 69(12):3535 – 3557, 2012. doi: 10.1175/JAS-D-12-0110.1. URL <https://journals.ametsoc.org/view/journals/atsc/69/12/jas-d-12-0110.1.xml>.
- James D. Doyle, Clark Amerault, Carolyn A. Reynolds, and P. Alex Reinecke. Initial condition sensitivity and predictability of a severe extratropical cyclone using a moist adjoint. *Monthly Weather Review*, 142(1):320 – 342, 2014. doi: 10.1175/MWR-D-13-00201.1. URL <https://journals.ametsoc.org/view/journals/mwre/142/1/mwr-d-13-00201.1.xml>.
- James D Doyle, Carolyn A Reynolds, and Clark Amerault. Adjoint sensitivity analysis of high-impact extratropical cyclones. *Monthly Weather Review*, 147(12):4511–4532, 2019.
- Daniel J. Lloveras, James D. Doyle, and Dale R. Durran. Can observation targeting be a wild goose chase? an adjoint-sensitivity study of a u.s. east coast cyclone forecast bust. *Journal of the Atmospheric Sciences*, 82(2):343 – 360, 2025. doi: 10.1175/JAS-D-24-0044.1. URL <https://journals.ametsoc.org/view/journals/atsc/82/2/JAS-D-24-0044.1.xml>.

- Jorge Baño-Medina, Agniv Sengupta, James D. Doyle, Carolyn A. Reynolds, Duncan Watson-Parris, and Luca Delle Monache. Are ai weather models learning atmospheric physics? a sensitivity analysis of cyclone xynthia. *npj Climate and Atmospheric Science*, 8(1):92, 2025. doi: 10.1038/s41612-025-00949-6. URL <https://doi.org/10.1038/s41612-025-00949-6>.
- Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.
- Vikki Thompson, Alan T Kennedy-Asser, Emily Vosper, YT Eunice Lo, Chris Huntingford, Oliver Andrews, Matthew Collins, Gabrielle C Hegerl, and Dann Mitchell. The 2021 western north america heat wave among the most extreme events ever recorded globally. *Science Advances*, 8(18):eabm6860, 2022.
- Nicholas J Leach, Christopher D Roberts, Matthias Aengenheyster, Daniel Heathcote, Dann M Mitchell, Vikki Thompson, Tim Palmer, Antje Weisheimer, and Myles R Allen. Heatwave attribution based on reliable operational weather forecasts. *Nature Communications*, 15(1):4530, 2024.
- Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.
- Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, pages 1–6, 2023.
- Matthias Karlbauer, Nathaniel Cresswell-Clay, Dale R. Durran, Raul A. Moreno, Thorsten Kurth, Boris Bonev, Noah Brenowitz, and Martin V. Butz. Advancing parsimonious deep learning weather prediction using the healpix mesh. *Journal of Advances in Modeling Earth Systems*, 16(8):e2023MS004021, 2024. doi: <https://doi.org/10.1029/2023MS004021>. URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023MS004021>.
- Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russel, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, et al. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *arXiv preprint arXiv:2308.15560*, 2023.
- Xiaofan Li, Zeng-Zhen Hu, Yu-heng Tseng, Yunyun Liu, and Ping Liang. A historical perspective of the la niña event in 2020/2021. *Journal of Geophysical Research: Atmospheres*, 127(7), 2022. doi: 10.1029/2021jd035546.
- Roger Daley and Thomas Mayer. Estimates of global analysis error from the global weather experiment observational network. *Monthly Weather Review*, 114(9):1642 – 1653, 1986. doi: 10.1175/1520-0493(1986)114<1642:EOGAEF>2.0.CO;2. URL [https://journals.ametsoc.org/view/journals/mwre/114/9/1520-0493\\_1986\\_114\\_1642\\_eogaef\\_2\\_0\\_co\\_2.xml](https://journals.ametsoc.org/view/journals/mwre/114/9/1520-0493_1986_114_1642_eogaef_2_0_co_2.xml).
- Gregory J. Hakim. Vertical structure of midlatitude analysis and forecast errors. *Monthly Weather Review*, 133(3):567 – 578, 2005. doi: 10.1175/MWR-2882.1. URL <https://journals.ametsoc.org/view/journals/mwre/133/3/mwr-2882.1.xml>.
- Marcelo Peña and Zoltan Toth. Estimation of analysis and forecast error variances. *Tellus A: Dynamic Meteorology and Oceanography*, 66(1):21767, 2014. doi: 10.3402/tellusa.v66.21767. URL <https://doi.org/10.3402/tellusa.v66.21767>.
- Zongheng Li, Jun Peng, Lifeng Zhang, and Jiping Guan. Exploring the differences in kinetic energy spectra between the ncep fnl and era5 datasets. *Journal of the Atmospheric Sciences*, 81(2):363–380, February 2024. doi: 10.1175/JAS-D-23-0043.1. URL <https://doi.org/10.1175/JAS-D-23-0043.1>.
- Peter Bauer. What if? numerical weather prediction at the crossroads. *Journal of the European Meteorological Society*, 1:100002, 2024. ISSN 2950-6301. doi: <https://doi.org/10.1016/j.jemets.2024.100002>. URL <https://www.sciencedirect.com/science/article/pii/S2950630124000024>.
- Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014. URL <https://arxiv.org/abs/1312.6199>.
- Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. *CVPR*, 11 2016.
- Pedro Tabacof and Eduardo Valle. Exploring the space of adversarial images. *CoRR*, abs/1510.05328, 2015. URL <http://arxiv.org/abs/1510.05328>.
- Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. URL <https://arxiv.org/abs/1412.6572>.

- Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features, 2019. URL <https://arxiv.org/abs/1905.02175>.
- Jacob Mitchell Springer, Melanie Mitchell, and Garrett T. Kenyon. Adversarial perturbations are not so weird: Entanglement of robust and non-robust features in neural network classifiers. *ArXiv*, abs/2102.05110, 2021. URL <https://api.semanticscholar.org/CorpusID:231861678>.
- Noah D. Brenowitz, Yair Cohen, Jaideep Pathak, Ankur Mahesh, Boris Boney, Thorsten Kurth, Dale R. Durran, Peter Harrington, and Michael S. Pritchard. A practical probabilistic benchmark for ai weather models. *arXiv preprint arXiv:2401.15305*, 2401.15305v1, jan 2024. URL <https://arxiv.org/abs/2401.15305>. License: CC BY 4.0.
- Andrew J. Charlton-Perez, Helen F. Dacre, Simon Driscoll, Suzanne L. Gray, Ben Harvey, Natalie J. Harvey, Kieran M. R. Hunt, Robert W. Lee, Ranjini Swaminathan, Remy Vandaele, and Ambrogio Volonté. 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, apr 2024. ISSN 2397-3722. doi: 10.1038/s41612-024-00638-w. URL <https://doi.org/10.1038/s41612-024-00638-w>.
- Massimo Bonavita. On some limitations of current machine learning weather prediction models. *Geophysical Research Letters*, 51(12):e2023GL107377, 2024. doi: <https://doi.org/10.1029/2023GL107377>. URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023GL107377>. e2023GL107377 2023GL107377.
- Nathaniel Cresswell-Clay, Bowen Liu, Dale Durran, Zihui Liu, Zachary I. Espinosa, Raul Moreno, and Matthias Karlbauer. A deep learning earth system model for efficient simulation of the observed climate, 2025. URL <https://arxiv.org/abs/2409.16247>.
- Syed Zahid Husain, Leo Separovic, Jean-François Caron, Rabah Aider, Mark Buehner, Stéphane Chamberland, Ervig Lapalme, Ron McTaggart-Cowan, Christopher Subich, Paul A. Vaillancourt, Jing Yang, and Ayrton Zadra. Leveraging data-driven weather models for improving numerical weather prediction skill through large-scale spectral nudging, 2024. URL <https://arxiv.org/abs/2407.06100>.
- H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellán, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R.J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J-N. Thépaut. Complete era5 from 1940: Fifth generation of ecmwf atmospheric reanalyses of the global climate. [Dataset]. Copernicus Climate Change Service (C3S) Data Store (CDS), 2017. DOI: 10.24381/cds.143582cf.
- Ronald M Errico. What is an adjoint model? *Bulletin of the American Meteorological Society*, 78(11):2577–2592, 1997.
- James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL <http://github.com/google/jax>.
- Carlos A. L. Pires, Robert Vautard, and O. Talagrand. On extending the limits of variational assimilation in nonlinear chaotic systems. *Tellus A*, 48:96–121, 1996. URL <https://api.semanticscholar.org/CorpusID:122300156>.
- Kyle Swanson, Robert Vautard, and Carlos Pires. Four-dimensional variational assimilation and predictability in a quasi-geostrophic model. *Tellus A: Dynamic Meteorology and Oceanography*, Jan 1998. doi: 10.3402/tellusa.v50i4.14540.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2017. URL <https://arxiv.org/abs/1412.6980>.
- Daniel S. Wilks. *Statistical Methods in the Atmospheric Sciences*. Academic Press, 3 edition, 2011. ISBN 9780123850225.