Check for updates

Geophysical Research Letters

RESEARCH LETTER
10.1029/2024GL110651

Special Collection:
Advancing Interpretable AI/ML Methods for Deeper Insights and Mechanistic Understanding in Earth Sciences: Beyond Predictive Capabilities

Predictability Limit of the 2021 Pacific Northwest Heatwave From Deep-Learning Sensitivity Analysis

P. Trent Vonich[1,2] and **Gregory J. Hakim**[1]

[1]Department of Atmospheric Sciences, University of Washington, Seattle, WA, USA, [2]Air Force Institute of Technology, Wright-Patterson AFB, OH, USA

Abstract The traditional method for estimating weather forecast sensitivity to initial conditions uses adjoint models, which are limited to short lead times due to linearization around a control forecast. The advent of deep-learning frameworks enables a new approach using backpropagation and gradient descent to iteratively optimize initial conditions, minimizing forecast errors. We apply this approach to the June 2021 Pacific Northwest heatwave using the GraphCast model, yielding over 90% reduction in 10-day forecast errors over the Pacific Northwest. Similar improvements are found for Pangu-Weather model forecasts initialized with the GraphCast-derived optimal, suggesting that model error is an unimportant part of the perturbations. Eliminating small scales from the perturbations also yields similar forecast improvements. Extending the length of the optimization window, we find forecast improvement to about 23 days, suggesting atmospheric predictability at the upper end of recent estimates.

Plain Language Summary This study examines a deep-learning approach to understanding how small changes to initial conditions impact weather forecasts. Traditionally, a linear approach known as the adjoint method has been used to determine the sensitivity of forecasts to initial conditions. We leverage recent advancements in machine learning to find optimal initial conditions using the backpropagation method within deep-learning frameworks. This approach iteratively searches for initial conditions that produce the best forecasts. We apply this method to GraphCast forecasts of the June 2021 Pacific Northwest extreme heatwave. We find that small changes to the initial conditions yield nearly perfect 10-day weather forecasts of the heatwave in both the GraphCast and the Pangu-Weather models. This research suggests that a significant increase in forecast skill may be possible from existing observations through better estimates of initial conditions.

Key Points:

- We use nonlinear gradient descent to optimize initial conditions for weather forecasting with machine learning models
- Application to the Pacific Northwest June 2021 heatwave reduces 10-day forecast error by over 90 percent
- Forecast improvements are not sensitive to the forecast model, and derive mainly from analysis errors on synoptic and larger scales

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

P. T. Vonich,
tvonich@uw.edu

Citation:

Vonich, P. T., & Hakim, G. J. (2024). Predictability limit of the 2021 Pacific Northwest heatwave from deep-learning sensitivity analysis. *Geophysical Research Letters*, *51*, e2024GL110651. https://doi.org/10.1029/2024GL110651

Received 20 JUN 2024
Accepted 24 AUG 2024

Author Contributions:

Conceptualization: P. Trent Vonich, Gregory J. Hakim
Data curation: P. Trent Vonich, Gregory J. Hakim
Formal analysis: P. Trent Vonich, Gregory J. Hakim
Funding acquisition: Gregory J. Hakim

© 2024 The Author(s). This article has been contributed to by U.S. Government employees and their work is in the public domain in the USA.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

1. Introduction

Atmospheric predictability is limited by sensitivity to the initial conditions (e.g., Bauer et al., 2015; Lorenz, 1996), and analysis of operational forecasts suggest this limit is currently around 10 days (Zhang et al., 2019), but may be extended using ensemble prediction to 16–23 days (e.g., Buizza & Leutbecher, 2015). For any individual case, forecast sensitivity can be quantified by defining a forecast metric to measure errors, and then computing the relationship between the metric and the initial conditions. Using a metric based on forecast error, for example, one can compute the changes to the initial condition that reduce forecast error. The most common approach to this problem employs adjoint models (e.g., Doyle et al., 2019; Errico, 1997; Langland et al., 1995), which linearize perturbations about a control forecast, limiting the lead time over which the sensitivity can be determined to less than about 5 days. A similar approach is available using automatic differentiation in deep-learning frameworks to identify the sensitivity of outputs to inputs (Gagne et al., 2019; McGovern et al., 2019; Toms et al., 2020). There are several advantages to the deep-learning approach, including a massive increase in computational efficiency relative to physical models, which enables deep searches for optimal initial conditions using nonlinear gradient descent. We apply this method to the extreme heatwave of June 2021 over the Pacific Northwest (PNW) region of North America using the GraphCast (Lam et al., 2023) deep learning model and the JAX framework (Bradbury et al., 2018) for computing sensitivity using automatic differentiation, with backpropagation and gradient descent optimization to identify the initial conditions that produce the best 10-day forecast of the heatwave.

There are several aspects of the deep-learning method that distinguish it from adjoint methods. In the adjoint method, the error in the forecast is propagated backward in time to the initial conditions using the adjoint version of the forecast model. The adjoint model is the transpose of the linearized version of the forecast model and

19448007, 2024, 19, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024GL110651, Wiley Online Library on [03/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Investigation: P. Trent Vonich, Gregory J. Hakim
Methodology: P. Trent Vonich
Project administration: P. Trent Vonich, Gregory J. Hakim
Supervision: Gregory J. Hakim
Validation: P. Trent Vonich, Gregory J. Hakim
Visualization: P. Trent Vonich, Gregory J. Hakim
Writing – original draft: P. Trent Vonich, Gregory J. Hakim
Writing – review & editing: P. Trent Vonich, Gregory J. Hakim

involves making choices for how to handle on/off processes. It also involves explicit coding, which is tedious and expensive. In contrast, in deep learning frameworks the models consist of layers with linear operations and analytic nonlinear activation functions between layers. As a consequence, derivatives simply apply the chain rule on known linear and nonlinear functions, which is well suited for automatic differentiation. Backpropagation of gradients is used to train deep-learning models by adjusting model weight parameters to better fit the output to training data. Here we hold the model weights constant and backpropagate forecast error to adjust the initial condition in order to better fit the verification data in the forecast, similar in spirit to Toms et al. (2020). Another important difference with the deep-learning approach is optimization, which combines the forward (nonlinear) pass and backpropagation of gradients on the output layer. The fact that the forecast model is extremely computationally inexpensive is essential to this approach. Adjoint approaches typically use a single pass due to the high cost of running a fully nonlinear physics model, and bottlenecks associated with linearization. Nevertheless, optimization strategies similar in spirit to the backpropagation approach have been proposed using the adjoint technique with physics models (e.g., Mu, 2000; Mu et al., 2003).

During the past 2 years a number of deep-learning models have emerged with forecast skill comparable to operational models (Rasp et al., 2023). These models are typically trained on the ERA5 reanalysis data set (Hersbach et al., 2020) and also appear to have encoded aspects of the basic laws of physics into their parameters as suggested by the idealized dynamical experiments of Hakim and Masanam (2024). However, this generation of models is also lacking power in forecasting smaller scales (Bonavita, 2023) and exhibit different forecast error growth from physics models at small amplitude (Selz & Craig, 2023). Because deep-learning models are trained to capture predictable signals, they inevitably filter out unpredictable signals, which leads to smoothing somewhat analogous to ensemble averaging (Bonavita, 2023). Although our sensitivity method will inherit these deficiencies, it is general and future improved models will likely overcome current deep-learning model weaknesses.

We select the PNW June 2021 heatwave for an example illustration of the method for two reasons. First, it is an exceptional heatwave, exceeded globally by only five other events since 1960 (Thompson et al., 2022), and is not included in the training data for the models we use. The event was due to an exceptional blocking anticyclone in the upper troposphere that was affected by latent heating from water vapor transported across the North Pacific ocean the preceding week (Lin et al., 2022; Oertel et al., 2023; Schumacher et al., 2022). Seasonal timing near the solstice and anomalously dry soil conditions (Conrick & Mass, 2023) also appear to play a contributing role. As such, this case provides a good test of deep-learning models in simulating an unseen extreme event with a mixture of contributing physical processes. Second, operational forecasts of the event from the European Center for Medium Range Weather Forecasts (ECMWF) Integrated Forecast System (IFS) were very skillful in predicting the event at least 7 days in advance (Emerton et al., 2022; Lin et al., 2022), although the amplitude of the event was underpredicted. At longer lead times, out to about 11 days, the IFS ensemble mean captured the event, but with much weaker amplitude; however, one ensemble member produced an excellent forecast (Leach et al., 2024). The fact that one ensemble member produced a very good forecast suggests that the ∼10-day forecasts of this event are potentially improvable by modifying the initial conditions, and we seek to determine to what extent our method can find an optimal initial condition that achieves this objective.

The remainder of the paper is organized as follows. In Section 2 we discuss the GraphCast model (Lam et al., 2023), which is used for forecast optimization, and the data used to initialize and verify forecasts. The method for forecast optimization is described in Section 3, along with the loss function we optimize using backpropagation and gradient descent. Results are presented in Section 4, for both GraphCast and the Pangu-Weather model (Bi et al., 2023) to test for the importance of model error in the optimal initial conditions. After presenting the optimized 10-day forecast for the PNW heatwave, we demonstrate the robustness of our method by optimizing forecasts for 10 different lead times using a different initialization time. A concluding discussion is provided in Section 5.

2. Model and Data

2.1. Graphcast

The version of the GraphCast model (Lam et al., 2023) used here is the "small" version, which predicts six atmospheric state variables (geopotential height, temperature, water vapor specific humidity, vertical velocity, and horizontal wind components) on 13 pressure levels, four surface variables (mean-sea-level pressure, 2m air

19448007, 2024, 19, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024GL110651, Wiley Online Library on [03/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

temperature, and 10m wind components), and 6-hr accumulated precipitation on a 1.0° grid. The model contains 36.7 million parameters and was trained on 1.0° ERA5 reanalysis data from 1979 to 2015. Inference consists of two atmospheric input states separated by six hours, and one output state six hours in the future. The output is fed autoregressively along with the preceding six-hour state as input to continue the forecast indefinitely. Here we only perturb the state variables, leaving specified variables such as the land-sea mask and top-of-atmosphere radiation fields unmodified.

2.2. Input Data

Our objective is to reduce forecast error as defined by the Graphcast loss function used in training (defined in Section 3). The main experiments pertain to optimizing the 10-day heatwave forecast initialized 00 UTC 20 June and valid 00Z 30 June 2021 since this is around the time that operational forecasts first began to resolve the event. Additional experiments are then presented for a wide range of forecast optimization lead times initialized 06 UTC 1 June 2021 to illustrate that the findings from the main experiments are not peculiar to that case.

3. Methods

3.1. Gradient-Based Input Optimization

Given input state $\mathbf{x}_i$ for iteration i ($i = 1$ represents ERA5 data), we compute an increment based on the gradient of the forecast loss, $\mathcal{L}(\mathbf{N}(\mathbf{x}_i))$ with respect to the inputs, where $\mathbf{N}$ represents GraphCast inference to the selected forecast optimization time:

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

The derivative in (Equation 1) involves mapping the gradient of the output back through the GraphCast neural network for every 6h time step. The gradient quantifies the sensitivity of the forecast loss to the input values as for adjoint sensitivity, and the step size η depends on the optimization algorithm. Toms et al. (2020) illustrate this approach for an El Niño Southern Oscillation (ENSO) classifier model, building on work from image classification tasks (Olah et al., 2017; Simonyan et al., 2013; Yosinski et al., 2015). Here we seek the inputs that produce the smallest forecast error over a chosen lead time as defined by the loss.

3.2. Optimization Process

The method used to produce a set of optimized inputs is as follows.

1. Given a set of inputs, produce a forecast at a chosen lead time. Here we initialize the first iteration with an ERA5 reanalysis state.
2. Calculate the forecast error using the loss function by verification against ERA5.
3. Calculate the gradient of the loss function with respect to the inputs using the JAX framework. JAX provides a facility for automatic differentiation, as well as GPU acceleration and dynamic code optimization (Bradbury et al., 2018).
4. Update the inputs using the Adam optimizer (Kingma & Ba, 2017) for gradient descent, applying the loss gradient as per Equation 1. For optimizations up to 14 days, we use the default parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a learning rate of 10^{-3}. For longer optimizations, we reduce the learning rate to 10^{-4} to handle the more complex gradient descent at extended lead times.
5. Repeat 1–4) until the loss plateaus or 100 iterations is reached. An example of the loss as a function of iteration is shown in Figure S1 in Supporting Information S1.

3.3. The Loss Function

We use the same scalar loss used to train Graphcast, which in essence is a weighted mean-squared-difference between the target output and the predicted output averaged across time, variable, and space. For predicted state $\hat{x}$ and verification state x, the loss is defined as

19448007, 2024, 19, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024GL110651, Wiley Online Library on [03/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

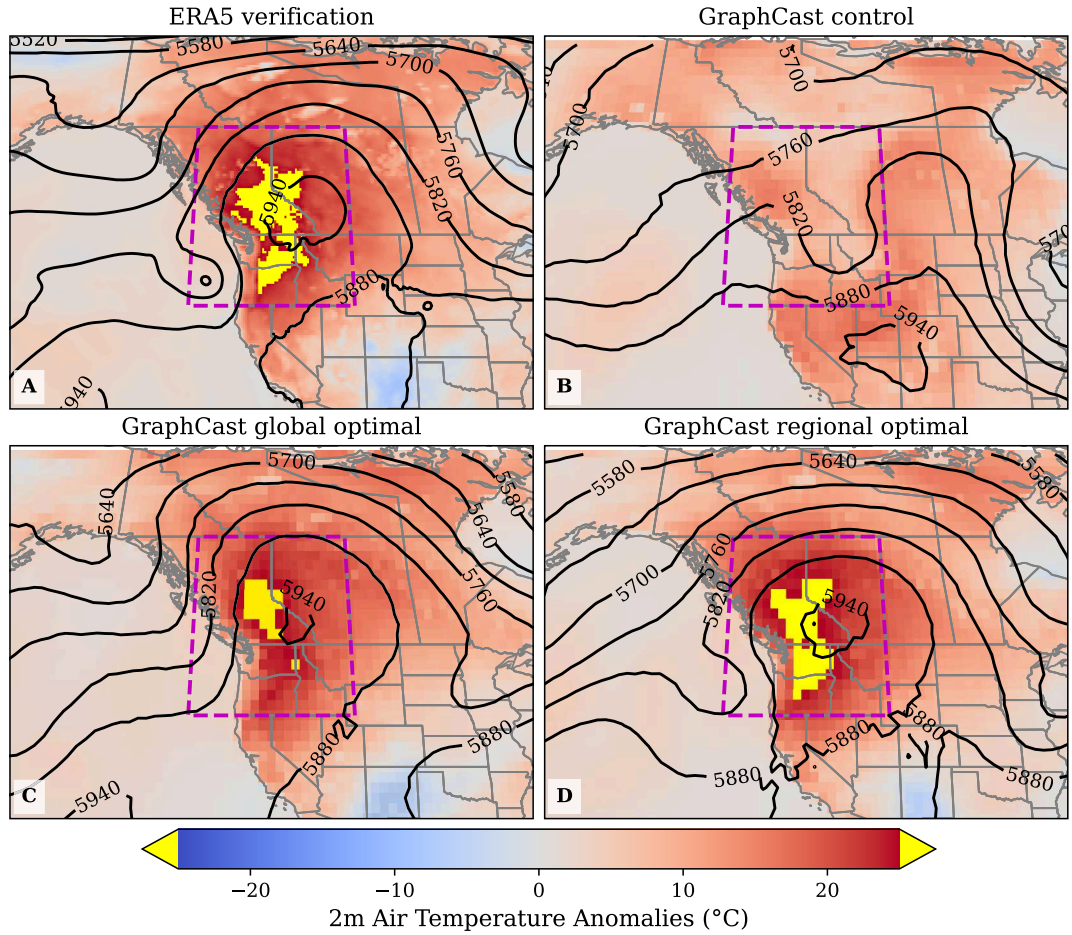

Figure 1. 500 hPa geopotential height (contours) and 2 m air temperature anomalies (colors) at 00 UTC 30 June 2021 for (a) ERA5 reanalysis, (b) GraphCast 10-day control forecast, (c) GraphCast 10-day forecast from globally optimal initial condition, and (d) GraphCast 10-day forecast from regionally optimal initial condition. All forecasts are initialized 00 UTC 20 June 2021. Anomalies are relative to ERA5 June climatology 1979–2020, with anomalies greater than 25°C shaded yellow.

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

Here, w is a 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. We refer the reader to Section 4.2 of the Materials and Methods section of Lam et al. (2023) for more details. We optimize for two spatial domains (i.e., G), one global, and the other specific to the PNW (42.0°N to 60.0°N, 130.0°W to 110.0°W). Finally, we note that, due to forecast error growth, the time average is heavily weighted toward times closer to the final (validation) time. Moreover, in figures shown subsequently, we plot the loss as a function of time without the time average in order to better reveal error changes in time.

4. Results

4.1. Ten-Day Optimized Heatwave Forecast

The target ERA5 verification field at 00 UTC 30 June shows conditions at the peak of the PNW heatwave (Figure 1a). At 500 hPa, a large ridge of high geopotential height with maximum values over 5940m is located over the region of greatest 2m temperature anomalies with values in excess of 25°C from the ERA5 June monthly

19448007, 2024, 19, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024GL110651, Wiley Online Library on [03/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

mean 1979–2020 (yellow shading). A region of low geopotential height is located west of the ridge, with a broad trough in the Gulf of Alaska, and a smaller trough west of the Oregon coast. The control 10-day forecast for GraphCast, initiated from the ERA5 analysis at 00 UTC 20 June, shows a trough over the PNW near the ridge in the verification field (Figure 1b). Control forecast 2m temperature anomalies in the region are slightly above normal, partly reflecting seasonality (30 June verification compared to June climatology), but also the fact that the trough is located in the middle of a larger-scale ridge over western North America. Note the difference in resolution between the GraphCast forecasts (1° as compared to ERA5 0.25°).

There are two results for GraphCast-optimized initial conditions: one that minimizes the loss function globally, and the other over the PNW region (denoted by magenta dashed lines in Figure 1). Even though the global-optimized initial condition does not specifically target the PNW, the forecast significantly improves upon the control in this region (Figure 1c). The trough over the PNW in the control is replaced by a ridge, similar to the verification field, and 2m air temperature anomalies in excess of 25°C are located in the same location as in the verification field, at the lower resolution of GraphCast. The regionally optimized solution produces a nearly perfect 10-day forecast over the PNW, down to the details of the ridge and trough at 500 hPa, and the amplitude and structure of the 2m temperature anomalies (Figure 1d). The evolution of the PNW loss as a function of time reveals that error for the globally optimized solution grows similarly to the control over the first 5 days, but then declines such that by day 8 the error is comparable to day 3 (Figure S2 in Supporting Information S1). Error then begins to increase again, but by 10 days the loss is ∼85% less than the control, with an implied exponential average doubling time of ∼2.5 days (estimate from the 10-day loss). The PNW-optimized solution shows little growth in error until day 8, and at day 10 the loss is over ∼90% less than the control, with an implied exponential average doubling time of ∼3.5 days (estimate from the 10-day loss).

While it is extraordinary that the initial conditions can be optimized to produce a nearly perfect 10-day forecast of the heatwave, two key questions remain: (a) to what extent do the initial conditions correct for model error rather than analysis error? and (b) are the initial condition changes practically realizable with an operational analysis and forecasting system? We address the first question by producing forecasts with a different model, Pangu-Weather, initialized with the GraphCast-optimized initial conditions. We address the second question by analyzing the magnitude of the initial-condition differences from the ERA5, and by filtering the optimal initial-condition differences from the control forecast to eliminate small spatial scales. Together, these experiments reveal whether the improvements are due to infinitesimal changes at small scales ("butterflies") or finite-amplitude changes at synoptic and larger scales, resolvable by the current observing system.

4.2. Pangu-Weather Forecast With GraphCast-Optimized Inputs

For the imperfect model experiments, we use the Pangu-Weather model (Bi et al., 2023), which has been shown to have similar forecast skill to GraphCast (Rasp et al., 2023). The Pangu-Weather model architecture is distinctly different from GraphCast, as is the method of inference; we highlight the differences relevant to our experiments. For inputs, the models share isobaric levels and variables with the exception of vertical velocity and precipitation, which are not included in Pangu-Weather. The horizontal spatial resolution of Pangu-Weather is 0.25° as compared to 1° for the small version of GraphCast used here. In order to map the GraphCast-optimal initial condition to the Pangu-Weather grid, we interpolate by projecting each variable in the initial condition onto spherical harmonics, pad the result at small scales with zeros to the Pangu-Weather resolution, and backtransform to the 0.25° latitude–longitude grid. For inference, GraphCast employs an autoregressive procedure using two time levels separated by 6 hr to predict the next 6-hr step, whereas Pangu-Weather employs an autoregressive procedure using a single time step to predict the next 24-hr step. We use only the GraphCast-optimized input at 00 UTC 20 June 2021 for Pangu-Weather, while GraphCast uses this and the additional optimized input at 18 UTC 19 June 2021. In summary, the modeling frameworks differ significantly in resolution (both spatially and temporally), input variables, network architecture, and inference procedure. Nevertheless, since both models are trained on the ERA5 data set, they may have learned similar relationships between input-output relationships, so further tests with physics models will be needed to assess the sensitivity of the optimized states to model error.

The Pangu-Weather control 10-day forecast shows a solution similar to the GraphCast control, with a 500 hPa trough over the PNW as compared to a ridge in the verification field (Figure 2a). The Pangu-Weather 10-day forecast initialized with the GraphCast-optimized initial conditions reveals a remarkable improvement over the control for both 500 hPa geopotential height and 2m air temperature anomalies (Figure 2b). Notable differences

10.1029/2024GL110651

19448007, 2024, 19, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024GL110651, Wiley Online Library on [03/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

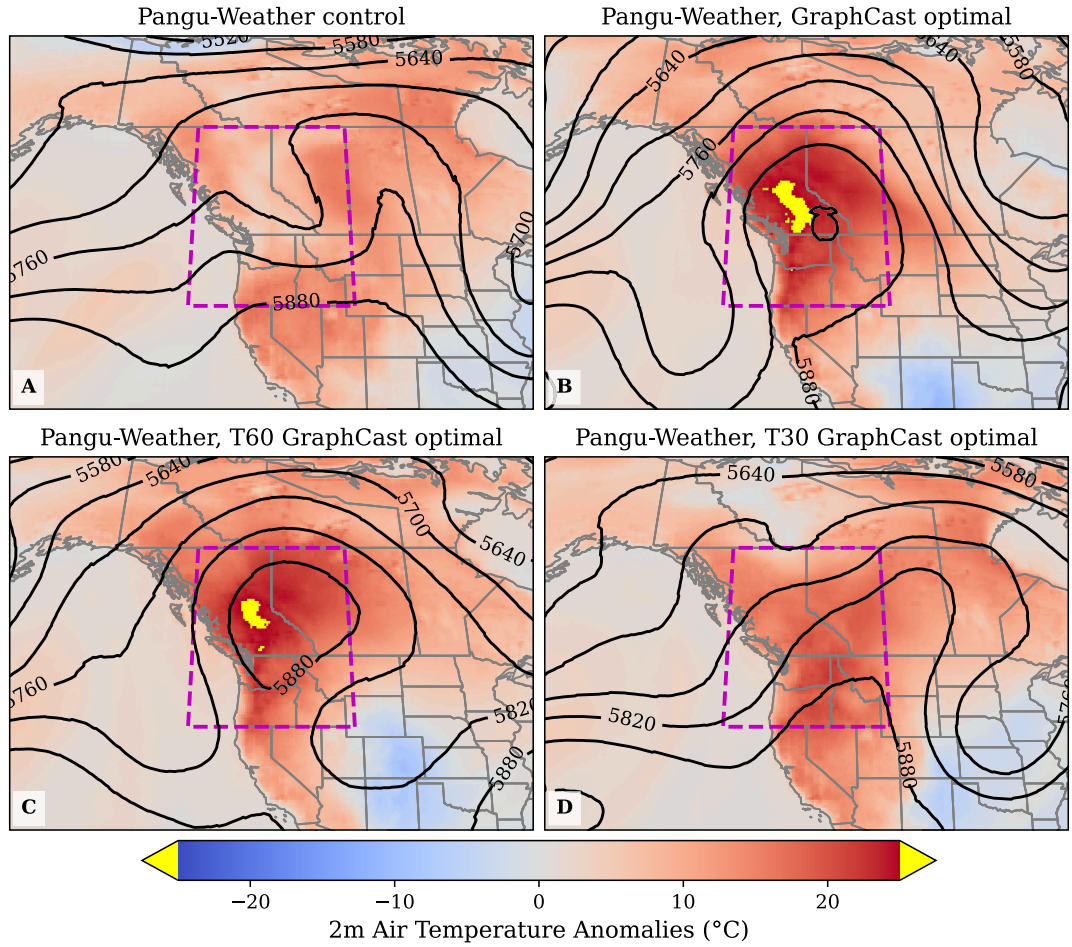

Figure 2. 500 hPa geopotential height (contours) and 2 m air temperature anomalies (colors) at 00 UTC 30 June 2021 for (a) Pangu-Weather 10-day control forecast, (b) Pangu-Weather 10-day forecast initialized with the GraphCast PNW-optimal initial condition (c) as in (b) but filtering the GraphCast optimal IC perturbations to T60 spherical harmonic truncation, and (d) as in (c) except for T30 GraphCast initial condition perturbations. All forecasts are initialized 00 UTC 20 June 2021. Anomalies are relative to ERA5 June climatology 1979–2020, with anomalies greater than 25°C shaded yellow.

from the verification fields include a slightly weaker ridge in the Pangu-Weather forecast, and 2m air temperature anomalies that are too small, especially over eastern Washington and Oregon, by approximately 3–5°C. This result suggests that deep-learning model error is not a significant contribution to the 10-day optimized initial condition, and that it largely captures analysis error in ERA5.

Turning to the second key question of whether the optimal initial condition is realizable by an operational forecast system, we first note that these changes are of modest magnitude, similar to observation and analysis errors (Table S1 in Supporting Information S1). Moreover, while the median values over the entire domain are small, they are not infinitesimal (i.e., not near floating point precision). The PNW-optimal perturbation maxima and medians are within 20% of globally optimized values and tend to be smaller (not shown). The geographical pattern of the PNW-optimized initial condition differences from ERA5 appear mostly as spatially white noise in the 500 hPa geopotential height field. Over 6 hr, structure evolves around the main extratropical waves and a wavenumber-2 pattern in the tropics (Figure S3a, S3b in Supporting Information S1). By 24 hr, the tropical patterns appear as wavenumber-1, and the amplitude of the differences is relatively larger in the extratropics (Figure S3c in Supporting Information S1). Differences from ERA5 for the first 24hr step of the Pangu-Weather solution from the GraphCast optimal show similar features in the extratropics, with less agreement in the tropics (Figure S3d in Supporting Information S1). In terms of the time-space evolution of forecast error changes in the 500 hPa geopotential height field, the optimal initial condition increases forecast error relative to ERA5 for lead times less than a day, followed by increasingly large improvements nearly everywhere as lead-time progresses (Figure S4 in

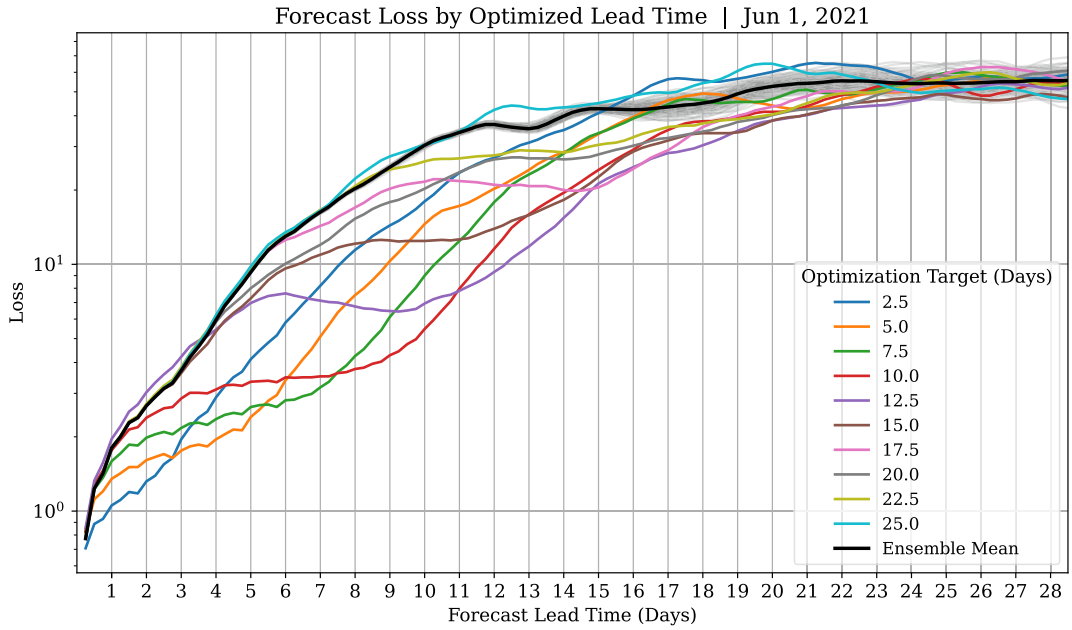

Figure 3. GraphCast instantaneous global loss (not time averaged) as a function of time starting 12 UTC 1 June 2021. A control ensemble forecast is shown in gray lines with the ensemble mean in black. All ensemble members have identical initial conditions given by the ERA5 analysis and diverge due to non-deterministic aspects of the GPU computing framework. Color lines show forecasts optimized for lead times ranging from 2.5 to 25 days.

Supporting Information S1). The rapid evolution of perturbations from noise in the initial conditions to structure appears related to the strong damping of small scales in the models, which implies a small affect on the loss function and therefore a small penalty during optimization. Moreover, the Adam optimizer also contributes to noise in the initial condition (not shown), and future work may consider other descent strategies such as conjugate gradient, which may not have this issue.

In order to determine the contribution from small scales in the 10-day forecast, we spectrally truncate the GraphCast PNW-optimized initial condition to T60 (~3° resolution), which eliminates over 99% of the degrees of freedom in spectral space. The Pangu-Weather forecast from this truncated initial condition is very similar to that for the full optimal initial condition (Figure 2c). Truncating further to T30 (~6° resolution) results in a weaker ridge and temperature anomalies, but still a significant improvement over the control forecast (Figure 2d). These experiments reveal that the important changes in the optimal initial condition are found on synoptic and planetary scales, and that the forecast differences are a smooth function of these truncated changes to the inputs. We conclude that these modest-amplitude, synoptic-scale analysis errors may be resolvable by the current observational network.

4.3. Globally Optimized Inputs at Multiple Target Lead Times

Given the dramatic increase in 10-day forecast skill from the optimal initial conditions for 00 UTC 30 June 2021, we now briefly address the generality of these results by optimizing for a wide range of forecast lead times, starting from a different initial time. Specifically, we consider 10 different optimization times from 2.5 to 25 days, now initialized at 12 UTC 1 June 2021. For all optimized initial conditions, we run the forecast for 28.5 days to 00 UTC 30 June in order to compare error growth. We compare the optimized forecasts against a control forecast initialized with the ERA5 reanalysis. The gray lines in Figure 3 show the forecast trajectories for 100 control forecasts initialized from identical ERA5 initial conditions, which evolve differently in time due to non-deterministic aspects of the GPU computing architecture. The control ensemble mean shows approximately exponential growth in forecast error from 1 to 7 days, followed by slower growth to saturation around 20 days (Figure 3).

Thin colored lines in Figure 3 show global forecast loss beginning at 12 UTC 1 June from initial conditions optimized for different lead times. For example, the 12.5-day optimized forecast (purple line) yields the most skillful forecast at 00 UTC 14 June 2021, and it shows very little error growth until that time. Although forecast

19448007, 2024, 19, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024GL110651, Wiley Online Library on [03/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19448007, 2024, 19, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024GL110651, Wiley Online Library on [03/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

error grows rapidly after the optimization time, the forecast from this initial condition improves upon the control ensemble out to ∼23 days. After a period of initial growth, forecast errors for 12.5-day and 17.5-day optimizations show error *decay* approaching the target time. This suggests that the optimization procedure has identified nearby initial conditions that nonlinearly evolve on the attractor near the true trajectory, but with small departures that allow the solution to remain close to the true state at the optimization time. The least effective optimization at long leads times is the 2.5-day optimal, which shows a delay in exponential growth, but is otherwise similar to the control.

The long-time limit of the optimization procedure appears to be around 22.5 days for the global loss, which improves upon the control even when considering an ensemble solution for the optimal initial condition (identical initial conditions; Figure S5 in Supporting Information S1, yellow lines). After 22.5 days, the optimal perturbations yield solutions no different than the control, and for completeness, we note that the long-time limit for PNW case is also around 22.5 days (not shown).

5. Discussion and Conclusion

Sensitivity to initial conditions is a hallmark of atmospheric predictability, ultimately limiting the accuracy of long-term forecasts. Traditionally this sensitivity is measured using adjoint models to estimate the features in the initial conditions that grow optimally at a chosen lead time as measured in a chosen norm. Since adjoint models involve linearizing about a control forecast, applications are typically limited to a few days. Moreover, the technical challenges of coding the adjoint model are time consuming and involve approximations for derivatives for on/off processes. Deep-learning models provide the opportunity for a new approach to atmospheric predictability since they are networks of linear models connected by known, differentiable, nonlinear activation functions. Importantly, the frameworks used to train these models offer tools to compute gradients with respect to not just model parameters, but state variables as well. In this paper, we define an algorithm that extends adjoint sensitivity to the fully nonlinear case to optimize initial conditions for forecast improvement at long lead times. This method's flexibility means that, in addition to the domain volumes considered here, input optimization may target specific variables, levels, timeframe, or location. We apply the method to the extreme heatwave over the Pacific Northwest in June 2021 using the GraphCast model and the JAX framework.

We find a ∼94% reduction in 10-day forecast error for the PNW heatwave over a control forecast initialized from the ERA5 reanalysis. In theory, the optimal initial conditions also encode corrections for model error, but similar forecast improvement from Pangu-Weather model forecasts starting from the GraphCast-optimized initial conditions suggests that the important perturbations are not specific to the forecast model. Truncating small scales from the optimal initial conditions reveals that the forecast improvements are due to finite-amplitude changes on synoptic and larger scales. The fact that the long-lead solutions vary smoothly in the truncation of the initial conditions suggests that, unlike physics models, the deep-learning models are not subject to noisy on/off processes (e.g., convective and microphysical parameterizations) that effectively add stochastic error to the forecast. Assessing the generality of this conjecture is an important direction for future research, especially in the context of interpreting differences in error growth between physics models and deep-learning models (Selz & Craig, 2023).

In addition to assessing the generality of the results presented here, future studies may also consider the joint optimization problem where both the model weights and the initial conditions are optimized. This could allow for flexible models that adapt optimally to a particular state and loss function, such as specific types of storms in a limited geographical area. Applications to data assimilation are also particularly compelling, as minimizing the loss from the misfit to observations (Hatfield et al., 2021) should allow for long-window 4DVAR relative to the short windows currently used (∼12 hr).

Data Availability Statement

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 the https://github.com/google-deepmind/graphcast (Lam et al., 2023) and https://github.com/198808xc/Pangu-Weather (Bi et al., 2023), respectively. Global and regional 10-day optimal initial conditions for 00 UTC 20 June 2021 are available at https://doi.org/10.5281/zenodo.13694959 (Vonich & Hakim, 2024).

Acknowledgments
We acknowledge high-performance computing support from the Casper cluster (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 Dr. Chris Snyder (NCAR) for conversations related to nonlinear singular vectors and initial-condition error growth, and Dr. Mu (Fudan University) for sharing his published papers on nonlinear adjoint methods. We thank Peter Dueben (ECMWF) and an anonymous referee for helpful comments that improved the clarity of the manuscript. PTV was funded by the Dr. Heather Wilson STEM Fellowship. GJH acknowledges support from NSF award 2202526 and Heising-Simons Foundation award 2023–4715.

References

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

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. [Software]. *Nature619*(7970), 1–6. https://doi.org/10.1038/s41586-023-06185-3

Bonavita, M. (2023). On the limitations of data-driven weather forecasting models. *arXiv preprint arXiv:2309.08473*.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., et al. (2018). JAX: Composable transformations of Python+NumPy programs. Retrieved from http://github.com/google/jax

Buizza, R., & Leutbecher, M. (2015). The forecast skill horizon. *Quarterly Journal of the Royal Meteorological Society*, *141*(693), 3366–3382. https://doi.org/10.1002/qj.2619

Conrick, R., & Mass, C. F. (2023). The influence of soil moisture on the historic 2021 Pacific Northwest heatwave. *Monthly Weather Review*, *151*(5), 1213–1228. https://doi.org/10.1175/mwr-d-22-0253.1

Doyle, J. D., Reynolds, C. A., & Amerault, C. (2019). Adjoint sensitivity analysis of high-impact extratropical cyclones. *Monthly Weather Review*, *147*(12), 4511–4532. https://doi.org/10.1175/mwr-d-19-0055.1

Emerton, R., Brimicombe, C., Magnusson, L., Roberts, C., Di Napoli, C., Cloke, H. L., & Pappenberger, F. (2022). Predicting the unprecedented: Forecasting the June 2021 Pacific Northwest heatwave. *Weather*, *77*(8), 272–279. https://doi.org/10.1002/wea.4257

Errico, R. M. (1997). What is an adjoint model? *Bulletin of the American Meteorological Society*, *78*(11), 2577–2592. https://doi.org/10.1175/1520-0477(1997)078<2577:wiaam>2.0.co;2

Gagne II, D. J., Haupt, S. E., Nychka, D. W., & Thompson, G. (2019). Interpretable deep learning for spatial analysis of severe hailstorms. *Monthly Weather Review*, *147*(8), 2827–2845. https://doi.org/10.1175/mwr-d-18-0316.1

Hakim, G. J., & Masanam, S. (2024). Dynamical tests of a deep-learning weather prediction model. *Artificial Intelligence for the Earth Systems*, *3*(3). https://doi.org/10.1175/aies-d-23-0090.1

Hatfield, S., Chantry, M., Dueben, P., Lopez, P., Geer, A., & Palmer, T. (2021). Building tangent-linear and adjoint models for data assimilation with neural networks. *Journal of Advances in Modeling Earth Systems*, *13*(9), e2021MS002521. https://doi.org/10.1029/2021ms002521

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2017). Complete era5 from 1940: Fifth generation of ecmwf atmospheric reanalyses of the global climate. [Dataset]. *Copernicus Climate Change Service (C3S) Data Store (CDS)*. https://doi.org/10.24381/cds.143582cf

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization. Retrieved from https://arxiv.org/abs/1412.6980

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

Langland, R. H., Elsberry, R. L., & Errico, R. M. (1995). Evaluation of physical processes in an idealized extratropical cyclone using adjoint sensitivity. *Quarterly Journal of the Royal Meteorological Society*, *121*(526), 1349–1386. https://doi.org/10.1002/qj.49712152608

Leach, N. J., Roberts, C. D., Aengenheyster, M., Heathcote, D., Mitchell, D. M., Thompson, V., et al. (2024). Heatwave attribution based on reliable operational weather forecasts. *Nature Communications*, *15*(1), 4530. https://doi.org/10.1038/s41467-024-48280-7

Lin, H., Mo, R., & Vitart, F. (2022). The 2021 western North American heatwave and its subseasonal predictions. *Geophysical Research Letters*, *49*(6), e2021GL097036. https://doi.org/10.1029/2021gl097036

Lorenz, E. N. (1996). Predictability: A problem partly solved. *Proc. seminar on predictability*, *1*.

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

Mu, M. (2000). Nonlinear singular vectors and nonlinear singular values. *Science in China - Series D: Earth Sciences*, *43*(4), 375–385. https://doi.org/10.1007/bf02959448

Mu, M., Duan, W., & Wang, B. (2003). Conditional nonlinear optimal perturbation and its applications. *Nonlinear Processes in Geophysics*, *10*(6), 493–501. https://doi.org/10.5194/npg-10-493-2003

Oertel, A., Pickl, M., Quinting, J., Hauser, S., Wandel, J., Magnusson, L., et al. (2023). Everything hits at once: How remote rainfall matters for the prediction of the 2021 north american heat wave. *Geophysical Research Letters*, *50*(3), e2022GL100958. https://doi.org/10.1029/2022gl100958

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, *2*(11), e7. https://doi.org/10.23915/distill.00007

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., & Russel, T. (2023). Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *arXiv preprint arXiv:2308.15560*.

Schumacher, D., Hauser, M., & Seneviratne, S. I. (2022). Drivers and mechanisms of the 2021 Pacific Northwest heatwave. *Earth's Future*, *10*(12), e2022EF002967. https://doi.org/10.1029/2022ef002967

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

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Thompson, V., Kennedy-Asser, A. T., Vosper, E., Lo, Y. E., Huntingford, C., Andrews, O., et al. (2022). The 2021 western North America heat wave among the most extreme events ever recorded globally. *Science Advances*, *8*(18), eabm6860. https://doi.org/10.1126/sciadv.abm6860

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically interpretable neural networks for the geosciences: Applications to Earth system variability. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2019MS002002. https://doi.org/10.1029/2019ms002002

Vonich, P., & Hakim, G. (2024). Predictability limit of the 2021 pacific Northwest Heatwave from deep-learning sensitivity analysis. [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.13694959

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.

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

19448007, 2024, 19, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024GL110651, Wiley Online Library on [03/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License