ADVANCING EARTH AND SPACE SCIENCES

Geophysical Research Letters

RESEARCH LETTER
10.1029/2025GL118886

Predictability of Storms in an Idealized Climate Revealed by Machine Learning

Wuqiushi Yao[1,2,3], Or Hadas[1], and Yohai Kaspi[1]

[1]Department of Earth and Planetary Science, Weizmann Institute of Science, Rehovot, Israel, [2]Key Laboratory of Earth System Numerical Modeling and Application, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China, [3]College of Earth and Planetary Sciences, University of Chinese Academy of Sciences, Beijing, China

Wuqiushi Yao and Or Hadas contributed equally to this work.

Key Points:

- We use CNN to predict storm tracks skillfully in an idealized GCM, providing a novel framework studying storm predictability
- Storm growth is less predictable than displacement, with baroclinicity and jet meanders revealing key source of uncertainties
- Quantitatively, jet meandering is associated with a doubling of the model's uncertainty sensitivity to jet structures

Supporting Information:

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

Correspondence to:

W. Yao and O. Hadas,
yaowuqiushi@gmail.com;
wuqiushi.yao@weizmann.ac.il;
or.hadas@weizmann.ac.il

Citation:

Yao, W., Hadas, O., & Kaspi, Y. (2026). Predictability of storms in an idealized climate revealed by machine learning. *Geophysical Research Letters*, *53*, e2025GL118886. https://doi.org/10.1029/2025GL118886

Received 20 AUG 2025
Accepted 22 DEC 2025

Author Contributions:

Conceptualization: Or Hadas, Yohai Kaspi
Data curation: Or Hadas, Yohai Kaspi
Formal analysis: Wuqiushi Yao, Or Hadas
Funding acquisition: Yohai Kaspi
Investigation: Wuqiushi Yao, Or Hadas
Methodology: Wuqiushi Yao, Or Hadas
Supervision: Or Hadas, Yohai Kaspi
Visualization: Wuqiushi Yao, Or Hadas, Yohai Kaspi

Abstract The midlatitude climate and weather are shaped by storms, yet the factors governing their predictability remain insufficiently understood. Here, we use a Convolutional Neural Network (CNN) to predict and quantify uncertainty in the intensity growth and trajectory of over 200,000 storms simulated with a 200-year aquaplanet GCM. This idealized framework provides a controlled climate background for isolating factors that govern predictability. Results show that storm intensity is less predictable than trajectory. Strong baroclinicity accelerates storm intensification and reduces its predictability, consistent with theory. Crucially, enhanced jet meanders further degrade forecast skill, revealing a synoptic source of uncertainty. Quantitatively, jet meandering over the storm center and eastern regions is associated with a doubling of the predicted uncertainty sensitivity in storm growth to the jet structure. These findings highlight the potential of machine learning for advancing understanding of predictability and its governing mechanisms.

Plain Language Summary Midlatitude storms are key drivers of weather and climate variability, yet their predictability remains limited. Using Convolutional Neural Network (CNN) to predict over 200,000 storms from a 200-year idealized climate simulation, we assess the factors controlling forecast accuracy. Our results reveal that storm intensity is significantly harder to predict than storm position, with errors growing fastest in regions of strong vertical wind shear. We further show that a more meandering upper-level jet stream reduces forecast skill by amplifying small initial uncertainties, a factor often overlooked in predictability studies. Incorporating Explainable AI, we pinpoint how storm forecast errors are particularly sensitive to subtle variations in jet structures. These findings highlight the potential of machine learning in understanding predictability of weather events.

1. Introduction

Cyclones play a crucial role in midlatitude atmospheric circulation, driving the meridional transport of heat, moisture, and momentum while shaping regional weather patterns and climate variability (e.g., Priestley & Catto, 2022; Tamarin & Kaspi, 2017). Understanding their predictability is therefore of critical importance, both in terms of forecasting individual storms and in identifying how the large-scale flow modulates their evolution. From a theoretical perspective, the chaotic nature of the atmosphere implies inherent limits to long-term predictability (Lorenz, 1963). This foundational insight, originally derived from idealized models, has driven decades of research into the sources and limits of atmospheric predictability. Previous studies have shown that storms embedded in stronger jet streams tend to intensify more rapidly and exhibit lower predictability (e.g., Doiteau et al., 2024; Froude et al., 2007; Pantillon et al., 2017). Over Europe, stronger storms have also been linked to greater ensemble spread and reduced forecast skill (Rupp et al., 2024). These findings align with the theoretical expectations that storms experiencing stronger wind shear (baroclinicity) will grow faster (Eady, 1949). Furthermore, Vallis (1983) directly connected increased baroclinicity to reduced predictability using an idealized model.

However, storms evolve within a wide range of climatic and synoptic environments (Hadas & Kaspi, 2026), which complicates quantifying predictability in the real atmosphere. This complexity obscures the fundamental relationships between background flow patterns and the predictability of individual storms. Idealized general circulation models (GCMs) provide a controlled and statistically stable climate framework that simplifies the dynamics while enabling long integrations and large storm samples (e.g., Chemke & Kaspi, 2015; Hadas & Kaspi, 2021; Tamarin & Kaspi, 2017; Walker & Schneider, 2006). The GCMs have

Writing – original draft: Wuqiushi Yao
Writing – review & editing: Or Hadas, Yohai Kaspi

therefore played an essential role in predictability studies across a range of climate conditions (Sheshadri et al., 2021).

Even in such simplified frameworks, assessing predictability and its response to perturbations of the initial state often requires large ensemble simulations, which are computationally expensive and complex (e.g., Coleman et al., 2024; DelSole, 2004; Emanuel & Zhang, 2016; Zhang & Tao, 2013). Machine learning offers an efficient alternative by directly learning the mapping between initial conditions and the associated spread of possible outcomes. Recent studies have demonstrated the potential of such probabilistic neural-network frameworks: Gordon and Barnes (2022) introduced a negative log-likelihood (NLL) loss that allows models to predict both the mean output (μ) and its associated uncertainty (σ), successfully identifying more predictable initial states, while Brettin et al. (2025) applied this approach to sea-level forecasts.

Therefore, combining machine learning with idealized GCMs is particularly fruitful: it allows efficient generation of large data sets and the development of accurate, uncertainty-aware models using relatively simple architectures compared with complex operational systems such as Pangu (Bi et al., 2023) and GraphCast (Price et al., 2025). The goal of this study is to uncover which aspects of the initial conditions control the forecast uncertainty of midlatitude storms. We train, validate, and test a CNN on 220,000 storm tracks from a 200-year idealized GCM simulation to predict storm displacement and vorticity growth. Section 2 describes the GCM setup, storm detection, and machine-learning methods. Section 3.1 assesses the overall predictability of storms, followed by analysis of how baroclinicity modulates predictability (Section 3.2) and how jet-stream meanders influence forecast uncertainty using an explainable AI technique (Section 3.3).

2. Data and Methods

2.1. Idealized GCM

Simulations are conducted using the Idealized Moist Spectral Atmospheric Model, with a T42 resolution for 200 years (Frierson et al., 2006). The model employs a spectral core that solves the primitive hydrostatic equations for an ideal-gas atmosphere (O'Gorman & Schneider, 2008). The simulations are carried out in an aquaplanet configuration, where the lower boundary is represented by a slab ocean with a prescribed heat capacity. Further details on the physical process parameterizations are provided in Tamarin and Kaspi (2017). The climatology of the large-scale circulation is shown in Figure S1 of Supporting Information S1, which highlights the subtropical jet (Figure S1a in Supporting Information S1), the eddy-driven jet, and the Hadley cells (Figure S1d in Supporting Information S1). The subtropical jet exhibits stronger vertical shear than the eddy-driven jet near 45°N, reflecting differences in baroclinicity between the two jets. Since the aquaplanet setup ensures hemispheric symmetry, data is aggregated from both hemispheres and shown as the northern hemisphere.

2.2. Tracking Algorithm

An objective feature-point identification and tracking technique (Hodges, 1995) is used to detect and track cyclones in the GCM. In this study, we identify cyclones between 20°N–60°N and between 20°S–60°S using the 838 hPa vorticity field, with their centers tracked every 6 hr. A minimum vorticity threshold of 10^{-5} s^{-1} is applied, and only cyclones that travel more than 1,000 km and persist for over 2 days are included in the analysis. About 220,000 cyclones are identified within a 200 years' run of GCM. Based on the tracking data, three properties of the storms are defined: the intensity change in terms of storm center vorticity (hereafter, growth, measured in s^{-1}), zonal and meridional displacements (Δx and Δy, respectively, measured in °).

To characterize the initial conditions under which storms form, we place a box around each cyclone's genesis location, extending ±30° meridionally and ±60° zonally. Then, the temperature, zonal wind, meridional wind, and specific humidity on 16 vertical levels, extending from 225 hPa to 963 hPa is recorded. The codes of storm tracking from GCMs and post processing for running the CNN are shown in Yao et al. (2025).

2.3. Machine Learning Model

In this study, we predict propagation and growth of storms (Δx, Δy, and growth) based on three-dimensional fields of the atmospheric state around the storm at genesis (see Section 2.2) using a machine learning model (Figure 1). The input and output variables are normalized by subtracting the ensemble mean across all storms and

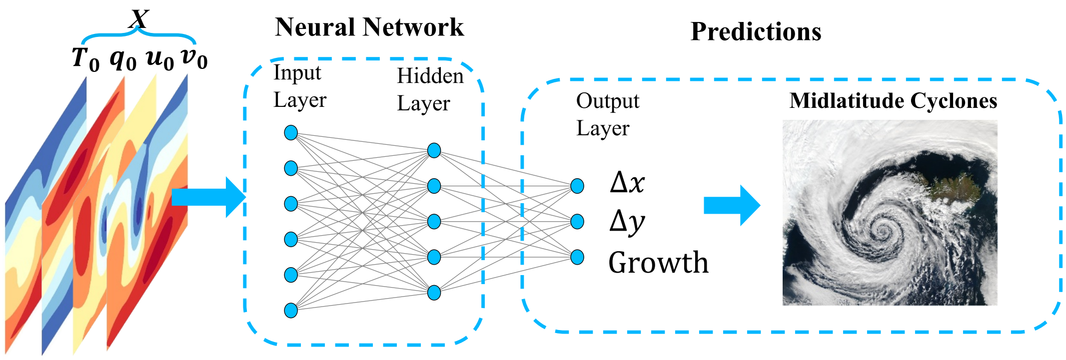

Figure 1. A schematic diagram illustrating the input and output structure of the neural network. The model takes atmospheric conditions around the storm at genesis as input and predicts storm growth, Δx and Δy (defined in Section 2.2) up to 42 hr ahead in 6-hr intervals.

dividing by the corresponding standard deviation at each grid point (see Table S1 in Supporting Information S1 for the mean and standard deviation of output variables).

The primary machine learning architecture employed in this study is a convolutional neural network (CNN). The CNN consists of six convolutional layers with hidden channel sizes of [92, 184, 184, 368, 368, 732], followed by three fully connected layers of sizes [512, 256, 16]. Each convolutional layer uses a kernel size of 3 with padding of 1. Except for the first layer (stride 1), a stride of 2 is applied, halving the spatial dimensions of the input maps at each layer. A dropout rate of 30% is applied after each convolutional layer to prevent overfitting, and ReLU is used as the activation function throughout. The model is trained using the Adam optimizer with a learning rate of 1×10^{-4}. For validation, we additionally implement a dense neural network (DNN) and a Linear Regression. Among these models, the CNN consistently achieves the best performance (Figure S2 in Supporting Information S1), while Linear Regression shows the worst performance and nearly no skill in predicting the storms. Thus, CNN serves as the basic model of this paper. This study utilizes 77,000 cyclones from the idealized GCM output for model training, 33,000 for validation, and 110,000 for testing. A large test set is employed to ensure the robustness of the analysis. All results shown are based on the "test" data set. The analysis focuses on the first 42-hr forecasts, during which the model exhibits its highest skill. All the codes regarding the model are included in Yao (2025).

In this study, Negative Log-Likelihood (NLL) loss function of a normal distribution (Barnes & Barnes, 2021; Guillaumin & Zanna, 2021) has been adopted, which allows models to jointly predict both the target variable and its associated uncertainty:

$$\mathcal{L}_\theta(\chi) = \frac{1}{2}\log 2\,\pi\sigma_\theta^2(\chi) + \frac{[\mathbf{f}(\chi) - \mu_\theta(\chi)]^2}{2\sigma_\theta^2(\chi)}, \tag{1}$$

where $\mathcal{L}$ represents the NLL loss function, and $\boldsymbol{\theta}$ denotes the learnable parameters of the model. $\boldsymbol{\chi}$ is the model input. The ground-truth output variable is denoted by $\mathbf{f}$, while the $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\sigma}_\theta^2$ are the predicted mean and variance.

In this framework, each output is treated as a Gaussian probability distribution: $\mathcal{L}_\theta(\chi) = -\log \boldsymbol{P}_\mathcal{N}\left(\mathbf{f}|\boldsymbol{\mu},\boldsymbol{\sigma}^2\right)$. The NLL loss acts to maximize the Gaussian likelihood of the observed target $\mathbf{f}$ under the model's predictive distribution $\boldsymbol{\mu},\boldsymbol{\sigma}^2$. Specifically, the formulation allows the model to reduce loss in multiple ways: by predicting a mean ($\boldsymbol{\mu}$) close to the target, by assigning a higher variance ($\boldsymbol{\sigma}^2$) to less predictable inputs, or both (Gordon & Barnes, 2022). As a result, high-error predictions are not penalized as long as they are accompanied by appropriately high predicted uncertainty, encouraging the model to learn and express initial-state-dependent predictability through the variable $\boldsymbol{\sigma}^2$. In practice, for the optimization, the $\mathcal{L}$ is averaged over all outputs and all lead times:

$$\mathcal{L}(\chi) = \frac{1}{3\cdot 7} \sum_{\substack{\mathrm{k}=1,2,3 \\ \text{(output variables)}}} \; \sum_{\substack{\mathrm{t}\in\{6,12,\ldots,42\}\ \mathrm{hr} \\ \text{(lead times)}}} \frac{1}{2}\left[\log\left(2\pi\,\sigma_{\mathrm{t,k}}^2(\chi)\right) + \frac{\left(\mathbf{f}_{\mathrm{t,k}}(\chi) - \boldsymbol{\mu}_{\mathrm{t,k}}(\chi)\right)^2}{\sigma_{\mathrm{t,k}}^2(\chi)}\right]. \tag{2}$$

where k denotes the three output variables: Δx, Δy, and growth, and t is the 7 lead time steps that we predict: 6, 12,…, 42 hr. Notably, the model input $\boldsymbol{\chi}$ is a four-dimensional matrix that contains the 3-D structure of 4 variables (see Section 2.2), giving each $\boldsymbol{\chi}$ a dimension of $\mathcal{R}^{4\times 43\times 21\times 16}$. For brevity, we omit the learnable parameters $\boldsymbol{\theta}$. Equation 2 is evaluated for every sample in the training set, and the loss is averaged over all the samples in the batch during model optimization. We evaluate the calibration of the uncertainty by grouping predictions with similar $\boldsymbol{\sigma}^2_{42\mathrm{h,k}}$ values and computing the realized MSE of the predicted means in each group. The observed MSE aligns with the predicted variance, indicating that the model's uncertainty estimates are well-calibrated (Figure S3, Text S1 in Supporting Information S1).

2.4. Explainable AI

In order to quantify the weight of each input variable in the forecast uncertainty, we apply sensitivity analysis (Simonyan et al., 2014), a gradient-based method widely used in machine learning to attribute model outputs to inputs. It allows to identify which variables dominate the prediction and to assess whether forecast errors or uncertainties are linked to these influential inputs. We define the sensitivity of the predicted mean of Gaussian as:

$$\mathcal{S}_{\mathrm{var,x,y,z,t,k}} = \mathrm{abs}\left[\frac{\partial \boldsymbol{\mu}_{\mathrm{t,k}}(\chi_{\mathrm{var,x,y,z}})}{\partial \chi_{\mathrm{var,x,y,z}}}\right], \tag{3}$$

where $\mathcal{S}$ measures the gradient of the output $\boldsymbol{\mu}$ to all the input variables, with the absolute value taken elementwise. The subscript "var" indicates the 4 input variables: zonal and meridional wind, temperature, and moisture. The subscripts x, y, z refer to the three spatial dimensions of the input fields, of sizes 43, 21, 16, respectively. The subscripts t, k are the output time step and output variable, respectively. Therefore, $\mathcal{S}$ is a matrix associated with both the shapes of the input and output fields: $\mathcal{R}^{\mathrm{input}\times\mathrm{output}} = \mathcal{R}^{(4\times 43\times 21\times 16)\times(7\times 3)}$. Since the NLL loss optimizes both the center and the spread of a Gaussian, we can also calculate the sensitivity of this uncertainty spread to the input:

$$\mathcal{E}_{\mathrm{var,x,y,z,t,k}} = \mathrm{abs}\left[\frac{\partial \boldsymbol{\sigma}^2_{\mathrm{t,k}}(\chi_{\mathrm{var,x,y,z}})}{\partial \chi_{\mathrm{var,x,y,z}}}\right], \tag{4}$$

where $\mathcal{E}$ quantifies how sensitive uncertainty of the output variables is to some small errors in the initial field $\boldsymbol{\chi}$. $\mathcal{E}$ is defined as the gradient of the output $\boldsymbol{\sigma}^2$ to all the input variables, with the absolute value taken elementwise. $\mathcal{E}$ has the same shape as $\mathcal{S}$. Together, $\mathcal{S}$ and $\mathcal{E}$ offer insight into how initial conditions contribute to the output and its uncertainty. Since in machine learning models all variables are normalized, the $\mathcal{S}$ and $\mathcal{E}$ are unitless, which allows comparison among variables.

3. Results

3.1. The Predictability of Cyclone Tracks

We first examine storm predictability as revealed by the CNN's forecasts applied to the aquaplanet GCM. For all variables: growth, Δx and Δy, the forecast error increases monotonically over time (Figure 2a; see Figure S4a in Supporting Information S1 for MSE expressed in physical units). In contrast, the R^2 peaks at the 18th hr rather than during the first few hours, likely because storm displacement and growth are too small to capture within the initial 12 hr (Table S1 in Supporting Information S1), making these early changes less relevant to the initial state (Figure S4b in Supporting Information S1). Storm growth remains the most challenging quantity to predict across different metrics, motivating our focus on growth in the following discussion.

To explore the predictability of growth, Figure 2b constructs a two-dimensional PDF of the predicted $\boldsymbol{\mu}_{42\mathrm{hr,growth}}$ and $\boldsymbol{\sigma}^2_{42\mathrm{hr,growth}}$ (abscissa and ordinate, respectively) of the predicted Gaussian distribution (Equation 1) for 42-hr growth. Variance reflects forecast uncertainty, with larger values indicating higher uncertainty. The 70% maximum density contour forms a tilted ellipse, while the contour connecting the peak density in each bin forms a distinct "V" shape, centered around the point of maximum density in the joint PDF of growth. These patterns indicate that rare events are inherently more difficult to predict as expected. In addition to this trend, considerable

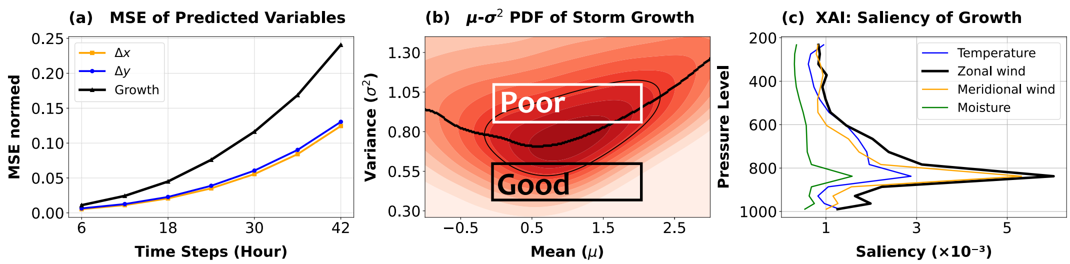

Figure 2. (a) Mean squared error (MSE) normalized by 42nd-hour variance of the CNN-predicted variables. MSE is formulated as $\mathrm{MSE}_{t,k} = \left(\mathbf{f}_{t,k} - \boldsymbol{\mu}_{t,k}\right)^2$, where t is lead time and k is the three predicted variables: Δx (yellow), Δy (blue), growth (black), respectively. MSE is computed and averaged over all the test-set samples. (b) Joint probability density function (PDF; shading) of the predicted $\boldsymbol{\mu}_{42\text{hr,growth}}$ and $\boldsymbol{\sigma}^2_{42\text{hr,growth}}$ (defined in Equation 1) from the machine learning models for the 42nd-hour growth. The thin black contour encloses 70% of the maximum density, and the thick contour indicates the peak density within each bin of predicted $\boldsymbol{\mu}_{42\text{hr,growth}}$ using bins of width $0.02 \times 10^{-5}\text{s}^{-1}$. The two boxes denote regions used to separate good and poor predictions. Contours are plotted at intervals of 0.03 from 0.0 to 0.3. (c) Vertical structure of the sensitivity (Equation 3) of 42-hr growth to all the inputs: $\mathcal{S}_{\text{var},z,42\text{hr,growth}} = \frac{1}{43\times 21}\sum_{x=1}^{43}\sum_{y=1}^{21}\mathcal{S}_{\text{var},x,y,z,42\text{hr,growth}}$. The $\mathcal{S}$ is computed for each test-set sample and then averaged over all samples.

variability in uncertainty remains even for a given growth rate. To test what drive this variability in predictability, two regions are selected that exhibit similar predicted growth (0–2.0 $\times$ 10^{-5} s^{-1}/42 hr but contrasting levels of uncertainty—one with low (good predictability) and the other with high (poor predictability) $\boldsymbol{\sigma}^2_{42\text{hr,growth}}$. Their implications for predictability will be examined in Section 3.3.

To uncover which input variables and physical processes contribute most to predictability, Figure 2c presents the sensitivity (Equation 3) of the 42-hr growth prediction to all the input variables. Sensitivity peaks at 838 hPa for all variables, as expected, since this is the storm-tracking level where environmental fields most directly influence vorticity dynamics. Among all variables, winds, which are directly linked to vorticity, contribute most to growth. Zonal wind slightly exceeds meridional wind due to its stronger association with the jet stream that guides storm motion. Mid-tropospheric temperature (500–750 hPa) surpasses meridional wind, reaching its maximum influence on growth within this layer. The results above demonstrate that the combination of machine learning and idealized GCM is able to make skillful and uncertainty-aware predictions, and identify which regions and variables contribute most to the predictability. Next, we would investigated what exact patterns are associated with predictability.

3.2. Baroclinicity and Predictability

In the aquaplanet simulation, the climate varies only with latitude due to the zonally symmetric boundary conditions. This feature allows us to directly relate storm predictability to latitude-dependent background conditions. Figure 3a shows that well-predicted storms preferentially reach the peak density near 44°N, whereas poorly predicted storms tend to cluster near 34°N. To explore the physical origin of this latitudinal dependence, we next examine the background climatology. In this idealized model, the most distinct and dynamically relevant feature of the climatology is the baroclinicity, which varies only with latitude. We quantify baroclinicity using the bulk vertical shear, defined as the difference in zonal wind between 275 hPa (model top layer) and 963 hPa (bottom layer). Using the Eady growth rate yields qualitatively similar results. The zonal-mean structure (Figure 3b) reveals maximum baroclinicity around 25°N, associated with the upper-level jet, and a weakening toward higher latitudes. Notably, this baroclinicity peak lies just south of the latitude band where poorly predicted storms concentrate (around 34°N), suggesting a possible dynamical link between baroclinicity and forecast uncertainty.

Consistent with Hadas and Kaspi (2025), storms embedded in more baroclinic regions exhibit faster growth (Figure 3c) and lower predictability, with the MSE of growth increasing with baroclinicity (Figure 3d). Interestingly, the error in Δy decreases with increasing baroclinicity, opposite to storm growth. Our interpretation is that stronger baroclinicity corresponds to enhanced upper-level potential-vorticity gradients, which improve the steering of storms by the large-scale flow (Tamarin & Kaspi, 2017), reducing sensitivity to small-scale, less predictable fluctuations. For Δx, this guiding effect seems to be offset by the chaotic background flow, leaving predictability nearly unchanged.

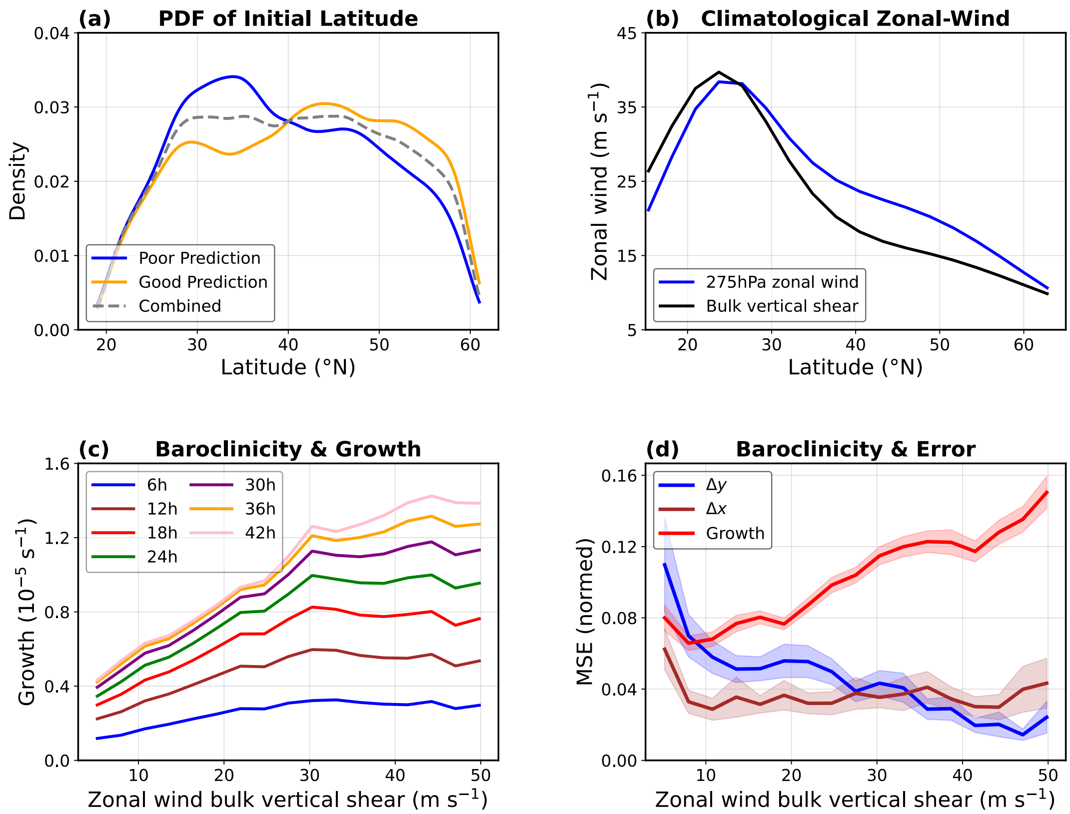

Figure 3. (a) PDF of the initial storm latitude for well predicted and poorly predicted storms (good and poor prediction, as defined in Figure 2b); (b) Climatological distribution of Zonal wind at 275 hPa (black) and bulk vertical shear (blue) in the idealized GCM; (c) The mean growth as a function of bulk vertical shear at the storm center, defined in Section 2.1, at different lead times; (d) The mean MSE of Δx, Δy and growth averaged over all time steps as a function of bulk vertical shear. Shading indicates ± 1 standard error, computed as the standard deviation within each bin divided by $\sqrt{N}$. MSE is calculated using variables normalized by the mean and standard deviation (Table S1 in Supporting Information S1).

Overall, the systematic alignment between latitude, vertical shear, and forecast skill demonstrates that storms forming in regions of stronger baroclinicity are less predictable in growth, but more predictable in Δy. Our results provide a clear link between baroclinicity and storm predictability in a medium-complexity aquaplanet model, consistent with findings from more realistic frameworks (Pirret et al., 2017).

3.3. Jet Meandering, Predictability and Error Growth Rate

We next seek to isolate the synoptic-scale conditions that govern forecast predictability, as recently shown in Hadas and Kaspi (2026). As discussed in Section 3.2, climatic conditions vary only with latitude in aquaplanet models. Therefore, separating storms according to latitude removes the climatic variability and leaves only the synoptic variability. To identify the dominant synoptic signal, Figure 4 presents composites of 275 hPa zonal wind for storms with "good" (Figures 4a and 4b) and "poor" (Figures 4d and 4e) predictability, separately for storms occurring north (Figures 4a and 4d) and south (Figures 4b and 4e) of 40°N. These structures have been confirmed to be insensitive to sample sizes (Figure S5 in Supporting Information S1).

In general, the 275 hPa zonal wind patterns differ markedly between the good- and poor-prediction groups, both south and north of 40°N (Figures 4a, 4b, 4d, and 4e). For storms forming south of 40°N, the good-prediction group (Figure 4a) exhibits a strong, coherent jet with a 36 m s^{-1} maximum at the storm center and relatively smooth meridional gradients ($\partial u/\partial y > 0$ south of the storm). In contrast, the poor-prediction group (Figure 4d) features a more intense core (>40 m s^{-1}) but displays strong zonal variability ($\partial u/\partial x$) east of the storm, with a pronounced meandering structure absent in the good-prediction counterpart. North of 40°N, both good- and poor-prediction

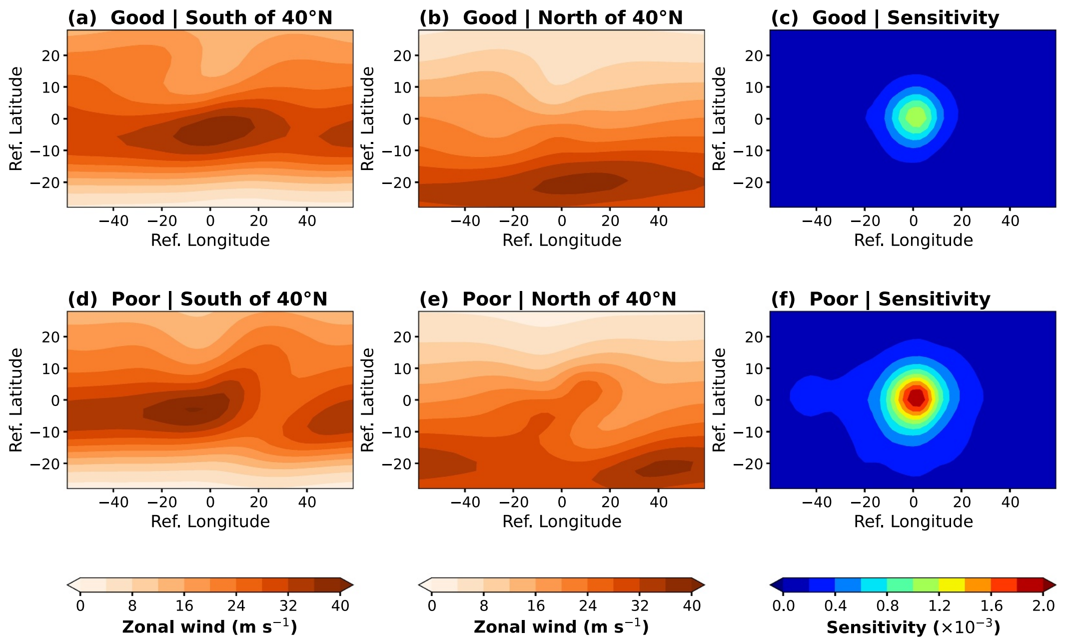

Figure 4. Composites of 275 hPa zonal wind fields at storm initialization for (a) well and (d) poorly predicted storms located south of 40°N; (b, e) are as (a, d), but for storms north of 40°N. The number of samples used for the composites are: (a) 6,824; (b) 8,677; (d) 7,406; and (e) 6,353. (c)Vertically integrated, horizontal distribution of the uncertainty saliency ($\mathcal{E}$, Equation 4) for the 42-hr growth-prediction sensitivity to the initial zonal-wind field. $\mathcal{E}_{\text{u-wind,x,y,42hr,growth}} = \frac{1}{16}\sum_{z=1}^{16}\mathcal{E}_{\text{u-wind,x,y,z,42hr,growth}}$, averaged for the well-predicted set defined in Figure 2; panel (f) Same as panel (c), but for the poorly predicted set.

storms are embedded in a jet regime characterized by decreasing wind speeds with latitude ($\partial u/\partial y < 0$), consistent with the climatological background (Figure 3c). The good-prediction group (Figure 4b) shows a relatively smooth jet with only weak meandering north of the storm. In contrast, the poor-prediction group (Figure 4e) reveals strong zonal asymmetry with sharp $\partial u/\partial x$ variations near and east of the storm center, while meridional gradients remain weak. A direct estimate of the relationship between $\partial u/\partial x$ and predictability (Figure S6 in Supporting Information S1) also supports that such a meandering structure increases forecast uncertainty of growth. Overall, the key distinction lies in the jet structure east of the storm center: poor-prediction storms are consistently associated with more meandering and zonally asymmetric upper-level winds, whereas well-predicted storms are embedded in smoother, more coherent jets. These structures likely modulate downstream uncertainty growth. Similar patterns are observed at lower levels (Figure S7 in Supporting Information S1).

In order to quantify how these structures are linked to predictability, here we evaluate the sensitivity of uncertainty to all the input variables (Equation 4). The results show that forecast uncertainty in 42-hr growth is most sensitive to variations in the zonal-wind structure (Figure S8 in Supporting Information S1), highlighting the dominant role of jet dynamics in storm predictability. The “good prediction” group exhibits roughly half the sensitivity of the “poor prediction” group to all the input variables, indicating a strong correspondence between reduced sensitivity of uncertainty (to input variables) and improved forecast skill. To probe the role of jet structure, we compare the sensitivity patterns of storms with well- and poorly predicted meandering zonal winds (Figures 4c and 4f). Because storms occurring south and north of 40°N display similar sensitivity characteristics, we present the composite of them. Both prediction groups show peak sensitivity near the storm center, but the poor-prediction group displays twice the sensitivity. This enhancement suggests that meandering structures east of the storm center substantially accelerate error growth. By contrast, a similar meander 10° north of the center in the good-prediction group (Figures 4a and 4b) contributes little to sensitivity, illustrating that such structures northward of the storm are irrelevant to the storm predictability (Figure 4c). Overall, these findings demonstrate that the spatial configuration of the jet, particularly downstream of the storm center, plays a critical role in shaping the predictability of storm intensity.

4. Conclusion and Discussion

This study investigates the predictability of midlatitude storms by applying a convolutional neural network (CNN) to forecast storm tracks in an aquaplanet GCM simulation. By integrating explainable AI with an idealized modeling framework, we find that:

1. Storm growth is less predictable than displacement, with large variability across storm samples. Zonal and meridional wind structures contribute most strongly to the predictions.
2. Stronger baroclinicity enhances storm growth while reducing its predictability, yet simultaneously improves the predictability of meridional displacement.
3. A more meandering jet is associated with a decreased storm growth predictability. Composite analysis links jet meandering to larger forecast uncertainty, while Explanable AI suggests that the model is sensitive in capturing such predictability changes.

These findings demonstrate that combining machine learning with traditional dynamical analysis can yield deeper insights into the mechanisms governing storm predictability. However, it remains uncertain whether the relationships identified in this idealized setting hold in the real atmosphere. Future work using reanalysis data and operational forecast models is essential to test the generalizability of these findings and evaluate their relevance for real-world weather prediction.

Global Research Statement

This research did not involve any external collaborators, local institutions, fieldwork, sampling, or community-based data collection. All data were generated via a 200-year aquaplanet experiment using the GFDL model, run on institutional high-performance computing systems. All individuals who met the AGU Publications authorship criteria were included as co-authors; those providing technical support are acknowledged in the Acknowledgments section. No formal permits, agreements with local authorities, or authorizations were required.

In line with AGU's mandate on "Inclusion in Global Research" and mindful of the TRUST Code's values of Fairness, Respect, Care, and Honesty, we ensured transparent attribution of contributions and open sharing of data and code. Co-authors jointly participated in the design, implementation, analysis, and interpretation stages of the experiment. The final manuscript was circulated to all co-authors, who verified the results and helped contextualize findings. Data and model scripts have been made publicly available in a trusted repository in accordance with the AGU Data and Software policy.

Overall, while no cross-regional or community-based collaboration was applicable, we confirm that ethical and scientific considerations have been met in full accordance with AGU policy and the TRUST Code values.

Conflict of Interest

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

Data Availability Statement

All datasets and code created or analyzed for this study are publicly archived in trusted repositories with persistent identifiers, enabling full transparency, reproducibility, and reuse. The data used in this study were generated using the publicly available Geophysical Fluid Dynamics Laboratory (GFDL) Idealized Moist Spectral Atmospheric Model (Frierson et al., 2006), which is fully documented and available at https://www.gfdl.noaa.gov/idealized-moist-spectral-atmospheric-model-quickstart/. Simulations were performed at T42 resolution for 200 years. Due to the large data volume, one representative simulation year, along with configuration files and post-processing scripts, has been archived in Zenodo (Yao et al., 2025). The full analysis and machine-learning code, including the CNN architecture, training algorithm, sensitivity-analysis tools, and figure-generation scripts, is archived at Zenodo (Yao, 2025).

Acknowledgments
W.Y acknowledges Huayu Chen for inspiring discussions regarding explainable AI. O.H acknowledges Laure Zanna for meaningful discussions. This research has been supported by the Azrieli Fellowship and the Israeli Science Foundation (Grant 996/20).

References

Barnes, E. A., & Barnes, R. J. (2021). Controlled Abstention neural networks for identifying skillful predictions for regression problems. *Journal of Advances in Modeling Earth Systems*, *13*(12), e2021MS002575. https://doi.org/10.1029/2021ms002575

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, *619*(7970), 533–538. https://doi.org/10.1038/s41586-023-06185-3

Brettin, A., Zanna, L., & Barnes, E. A. (2025). Uncertainty-permitting machine learning reveals sources of dynamic sea level predictability across daily to seasonal time scales. *Artificial Intelligence for the Earth Systems*, *4*(4), 250014. https://doi.org/10.1175/aies-d-25-0014.1

Chemke, R., & Kaspi, Y. (2015). The latitudinal dependence of atmospheric jet scales and macroturbulent energy cascades. *Journal of the Atmospheric Sciences*, *72*(10), 3891–3907. https://doi.org/10.1175/jas-d-15-0007.1

Coleman, A., Ancell, B., & Schwartz, C. (2024). Can we predict the predictability of high-impact weather events? *Monthly Weather Review*, *152*(11), 2483–2504. https://doi.org/10.1175/mwr-d-23-0293.1

DelSole, T. (2004). Predictability and information theory. Part I: Measures of predictability. *Journal of the Atmospheric Sciences*, *61*(20), 2425–2440. https://doi.org/10.1175/1520-0469(2004)061<2425:paitpi>2.0.co;2

Doiteau, B., Pantillon, F., Plu, M., Descamps, L., & Rieutord, T. (2024). Systematic evaluation of the predictability of different Mediterranean cyclone categories. *Weather and Climate Dynamics*, *5*(4), 1409–1427. https://doi.org/10.5194/wcd-5-1409-2024

Eady, E. T. (1949). Long waves and cyclone waves. *Tellus*, *1*(3), 33–52. https://doi.org/10.3402/tellusa.v1i3.8507

Emanuel, K., & Zhang, F. (2016). On the predictability and error sources of tropical cyclone intensity forecasts. *Journal of the Atmospheric Sciences*, *73*(9), 3739–3747. https://doi.org/10.1175/jas-d-16-0100.1

Frierson, D. M. W., Held, I. M., & Zurita-Gotor, P. (2006). A gray-radiation aquaplanet moist GCM. Part I: Static stability and eddy scale. *Journal of the Atmospheric Sciences*, *63*(10), 2548–2566. https://doi.org/10.1175/jas3753.1

Froude, L. S. R., Bengtsson, L., & Hodges, K. I. (2007). The predictability of extratropical storm tracks and the sensitivity of their prediction to the observing system. *Monthly Weather Review*, *135*(2), 315–333. https://doi.org/10.1175/mwr3274.1

Gordon, E. M., & Barnes, E. A. (2022). Incorporating uncertainty into a regression neural network enables identification of decadal state-dependent predictability in CESM2. *Geophysical Research Letters*, *49*(15), e2022GL098635. https://doi.org/10.1029/2022gl098635

Guillaumin, A. P., & Zanna, L. (2021). Stochastic-deep learning parameterization of ocean momentum forcing. *Journal of Advances in Modeling Earth Systems*, *13*(9), e2021MS002534. https://doi.org/10.1029/2021ms002534

Hadas, O., & Kaspi, Y. (2021). Suppression of baroclinic eddies by strong jets. *Journal of the Atmospheric Sciences*, *78*(8), 2445–2457. https://doi.org/10.1175/jas-d-20-0289.1

Hadas, O., & Kaspi, Y. (2025). A Lagrangian perspective on the growth of midlatitude storms. *AGU Advances*, *6*(3), e2024AV001555. https://doi.org/10.1029/2024AV001555

Hadas, O., & Kaspi, Y. (2026). Quantifying the influence of climate on storm activity using machine learning. *Geophysical Research Letters*, e2025GL118496. https://doi.org/10.1029/2025GL118496

Hodges, K. I. (1995). Feature tracking on the unit sphere. *Monthly Weather Review*, *123*(12), 3458–3465. https://doi.org/10.1175/1520-0493(1995)123<3458:ftotus>2.0.co;2

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

O'Gorman, P. A., & Schneider, T. (2008). The hydrological cycle over a wide range of climates simulated with an idealized GCM. *Journal of Climate*, *21*(15), 3815–3832. https://doi.org/10.1175/2007jcli2065.1

Pantillon, F., Knippertz, P., & Corsmeier, U. (2017). Revisiting the synoptic-scale predictability of severe European winter storms using ECMWF ensemble reforecasts. *Natural Hazards and Earth System Sciences*, *17*(10), 1795–1810. https://doi.org/10.5194/nhess-17-1795-2017

Pirret, J. S. R., Knippertz, P., & Trzeciak, T. M. (2017). Drivers for the deepening of severe European windstorms and their impacts on forecast quality. *Quarterly Journal of the Royal Meteorological Society*, *143*(702), 309–320. https://doi.org/10.1002/qj.2929

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

Priestley, M. D. K., & Catto, J. L. (2022). Future changes in the extratropical storm tracks and cyclone intensity, wind speed, and structure. *Weather and Climate Dynamics*, *3*(1), 337–360. https://doi.org/10.5194/wcd-3-337-2022

Rupp, P., Spaeth, J., Afargan-Gerstman, H., Büeler, D., Sprenger, M., & Birner, T. (2024). The impact of synoptic storm likelihood on European subseasonal forecast uncertainty and their modulation by the stratosphere. *Weather and Climate Dynamics*, *5*(4), 1287–1298. https://doi.org/10.5194/wcd-5-1287-2024

Sheshadri, A., Borrus, M., Yoder, M. R., & Robinson, T. A. (2021). Midlatitude error growth in atmospheric GCMs: The role of eddy growth rate. *Geophysical Research Letters*, *48*(24), e2021GL096126. https://doi.org/10.1029/2021gl096126

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proceedings of the workshop at the international conference on learning representations (ICLR)*. Banff.

Tamarin, T., & Kaspi, Y. (2017). The poleward shift of storm tracks under global warming: A Lagrangian perspective. *Geophysical Research Letters*, *44*(20), 10666–10674. https://doi.org/10.1002/2017gl073633

Vallis, G. K. (1983). On the predictability of quasi-geostrophic flow: The effects of beta and baroclinicity. *Journal of the Atmospheric Sciences*, *40*(1), 10–27. https://doi.org/10.1175/1520-0469(1983)040<0010:otpoqg>2.0.co;2

Walker, C. C., & Schneider, T. (2006). Eddy influences on Hadley circulations: Simulations with an idealized GCM. *Journal of the Atmospheric Sciences*, *63*(12), 3333–3350. https://doi.org/10.1175/jas3821.1

Yao, W. (2025). Code for studying storm predictability [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.16741969

Yao, W., Hadas, O., & Kaspi, Y. (2025). Idealized GCM dataset and post-processing results for studying storm predictability [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.16907625

Zhang, F., & Tao, D. (2013). Effects of vertical wind sheer on the predictability of tropical cyclones. *Journal of the Atmospheric Sciences*, *70*(3), 975–983. https://doi.org/10.1175/jas-d-12-0133.1