Using Neural Networks to Learn the Jet Stream Forced Response from Natural Variability

CHARLOTTE CONNOLLY,[a] ELIZABETH A. BARNES,[a] PEDRAM HASSANZADEH,[b,c] AND MIKE PRITCHARD[d,e]

[a] *Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado*
[b] *Department of Mechanical Engineering, Rice University, Houston, Texas*
[c] *Department of Earth, Environmental and Planetary Sciences, Rice University, Houston, Texas*
[d] *Department of Earth System Science, University of California, Irvine, Irvine, California*
[e] *NVIDIA Corporation, Santa Clara, California*

(Manuscript received 19 December 2022, in final form 14 March 2023, accepted 28 March 2023)

ABSTRACT: Two distinct features of anthropogenic climate change, warming in the tropical upper troposphere and warming at the Arctic surface, have competing effects on the midlatitude jet stream's latitudinal position, often referred to as a "tug-of-war." Studies that investigate the jet's response to these thermal forcings show that it is sensitive to model type, season, initial atmospheric conditions, and the shape and magnitude of the forcing. Much of this past work focuses on studying a simulation's response to external manipulation. In contrast, we explore the potential to train a convolutional neural network (CNN) on internal variability alone and then use it to examine possible nonlinear responses of the jet to tropospheric thermal forcing that more closely resemble anthropogenic climate change. Our approach leverages the idea behind the fluctuation–dissipation theorem, which relates the internal variability of a system to its forced response but so far has been only used to quantify linear responses. We train a CNN on data from a long control run of the CESM dry dynamical core and show that it is able to skillfully predict the nonlinear response of the jet to sustained external forcing. The trained CNN provides a quick method for exploring the jet stream sensitivity to a wide range of tropospheric temperature tendencies and, considering that this method can likely be applied to any model with a long control run, could be useful for early-stage experiment design.

KEYWORDS: Atmospheric circulation; Forcing; Nonlinear dynamics; Data science; Deep learning; Neural networks

1. Introduction

The eddy-driven jet stream drives much of the Northern Hemisphere midlatitude weather (e.g., Nakamura et al. 2004; Athanasiadis et al. 2010; Shaw et al. 2016; Madonna et al. 2017). Consequently, changes in the jet stream position and strength can result in enormous societal impact by impacting heat waves, droughts, and flooding events (Schubert et al. 2011; Coumou and Rahmstorf 2012; Bibi et al. 2020; Rousi et al. 2021, 2022), extreme weather across the midlatitudes (Mahlstein et al. 2012; Röthlisberger et al. 2016), hurricane tracks (Mattingly et al. 2015), and crop production (Kornhuber et al. 2020). Two robust features of anthropogenic climate change, warming in the upper troposphere of the tropics and warming at the surface of the Arctic, have been shown to independently force opposite responses in the mean jet location (e.g., Held 1993; Harvey et al. 2015; Stendel et al. 2021) These competing responses are driven by changes in the pole to equator temperature gradient (Blackport and Screen 2020; Stendel et al. 2021). Warming in the tropical upper troposphere drives a poleward shift in the mean jet location by increasing the upper-tropospheric temperature gradient, while simultaneously, warming at the Arctic surface drives an equatorward shift in the mean jet location by decreasing the surface temperature gradient (Butler et al. 2010; Screen et al. 2013; Chen et al. 2020; Stendel et al. 2021), The competing jet response stemming from these two thermal forcings is commonly referred to as the "tug-of-war" on the jet stream. Current consensus across climate models is that the upper-tropospheric warming wins out over the Arctic surface warming, causing a net poleward shift of the jet (Yin 2005; Swart and Fyfe 2012; Barnes and Polvani 2013; Harvey et al. 2015). However, there is still substantial disagreement over the magnitude of the jet response due to uncertainty in the strength and spatial extent of the regional heating anomalies (Grise and Polvani 2016).

Warming in both the tropical upper troposphere and Arctic surface are caused by distinctly different dynamical processes that determine the characteristics of the thermal anomalies. The tropical upper atmosphere warms more as a result of additional water vapor stored in the warmer tropical tropospheric air (i.e., a reduction in the moist adiabatic lapse rate; Sherwood and Nishant 2015). The enhanced Arctic warming, commonly referred to as Arctic Amplification, is occurring 3 times faster than elsewhere on the planet (Blunden and Arndt 2012; Druckenmiller et al. 2021) and is driven by multiple processes that include changes in poleward energy transport (Hwang and Frierson 2010; Graversen and Langen 2019), surface ice–albedo feedbacks (Manabe and Stouffer 1980; Dai et al. 2019), cloud feedbacks (Abbot and Tziperman 2008) and lapse-rate feedbacks (Pithan and Mauritsen 2014). To further

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/AIES-D-22-0094.s1.

Corresponding author: Charlotte Connolly, cconn@rams.colostate.edu

DOI: 10.1175/AIES-D-22-0094.1 e220094

increase the complexity of the processes driving the regional warming, the two thermal forcings likely do not act entirely independently. Research has shown that increased transient Rossby waves initiated in the tropics may drive increased heat transport into the high latitudes and as a result drive further warming in the mid- and upper troposphere of the Arctic (Lee 2014; Dunn-Sigouin et al. 2021). Uncertainties in the processes that contribute to the magnitude and shape of warming in the tropical upper troposphere and Arctic surface (Blackport and Screen 2020; Stendel et al. 2021), in turn, make it even more challenging to predict the magnitude of the jet response.

Despite the large body of work that investigates the response of the midlatitude jet under climate change, multiple challenges, such as the short observational record, isolating the jet's forced response from internal variability, and modeling ice and cloud feedbacks continue to make the question difficult to answer (Tjernström et al. 2008; Kattsov et al. 2010; Cohen et al. 2014; Pithan and Mauritsen 2014; Vihma 2014). The studies that have investigated the response of the jet to a thermal forcing have shown that the jet is sensitive to the shape, location, and magnitude of the thermal forcing (Butler et al. 2010), the season in which the forcing is imposed (McGraw and Barnes 2016), the current state of the atmosphere (i.e., position of the jet stream; Gerber et al. 2008; Barnes et al. 2010; Kidston and Gerber 2010; Garfinkel et al. 2013), and the climate models used for the study (Meehl et al. 2007; Barnes and Polvani 2013).

In an attempt to explore circulation sensitivities to a wide range of possible thermal forcings, Hassanzadeh and Kuang (2016b) used a control run from the GFDL dry dynamical core (Manabe et al. 1974) and employed the fluctuation–dissipation theorem (FDT) to compute the linear response function of the circulation to a number of external thermal and mechanical forcing. FDT relates the mean linear response of a nonlinear system to a forcing through a linear operator created from the internal variability of the system (e.g., Kraichnan 1959; Leith 1975; Marconi et al. 2008). With the ability to explore a forced response from internal variability, FDT has been proposed as a method to quickly estimate circulation sensitivities in climate models (Fuchs et al. 2015) and serve as a useful tool for planning expensive climate model experiments (Leith 1975). There have been encouraging results using FDT to explore the circulation response to thermal forcings in general circulation models (Gritsun and Branstator 2007) as well as more complex coupled climate models (Phipps 2010) to estimate the response to realistic sea surface thermal forcings (Fuchs et al. 2015).

In order for the linear operator of FDT to accurately predict the mean response to a forcing, the system must satisfy a number of conditions (Marconi et al. 2008). The first condition is that the system must be in equilibrium, because FDT assumes that small changes in the system's state (internal variability) has a recovery back to equilibrium that is similar to the system's response to a small perturbation (Kraichnan 1959; Leith 1975). The second is that the perturbation must be small enough so that the response is linear even though the system that the operator is created from is not necessarily linear (Leith 1975). Last, the probability density function of the system must be differentiable, and many applications of the FDT assume the system probability density function is Gaussian (Majda et al. 2005), though work has been done to make versions of FDT where the system can be quasi-Gaussian (Cooper and Haynes 2011). In theory, a system that satisfies these conditions can use FDT to compute the systems' linear response to a forcing, though there are practical challenges in applying FDT to high-dimensional systems, such as GCMs (Lutsko et al. 2015; Hassanzadeh and Kuang 2016b; Khodkar and Hassanzadeh 2018).

Instead of using FDT to relate a forcing to a response, this study uses a convolutional neural network (CNN) to learn the nonlinear relationship between a forcing and a response. Moreover, using a CNN in place of the linear operator removes the need to make some of the FDT assumptions (i.e., small forcing for a linear response, Gaussianity assumption). Training is performed on data from a long control run with the Community Earth System Model (CESM) dry dynamical core. Once trained, the CNN is used to explore the jet sensitivity to a variety of thermal forcings. Throughout this study, we evaluate the CNN's ability to quantify the CESM dry dynamical core's jet sensitivity, placing particular emphasis on the tug-of-war between the warming in the tropical upper troposphere and the Arctic surface. Training a neural network on internal variability alone and then using it to predict a forced response is, to our knowledge, a novel application of deep learning to climate analysis. Therefore, we assess the strengths and weaknesses of this approach in multiple ways (see section 3).

2. Methods

We train a CNN to predict the jet stream response to zonally averaged regional temperature perturbations. The goal is to investigate jet sensitivity to thermal features associated with anthropogenic climate change. The CNN, detailed in section 2b, is trained on a long control run from a dry dynamical core, which is documented to reproduce the majority of the Northern Hemisphere's jet response to heating perturbations along with simulating the correct sign of the jet shift (e.g., Mbengue and Schneider 2013; Hassanzadeh et al. 2014; McGraw and Barnes 2016; Baker et al. 2017). Once trained, the CNN's skill is examined by comparing it with additional baseline prediction methods and dry core experiments that include an imposed thermal forcing. Details on the dry dynamical core setup, the CNN architecture and training, additional baseline prediction methods, and additional dry core heating experiments are discussed in more detail in the following sections.

a. Training data

We use output from the CESM Eulerian spectral-transform dry dynamical core (Lauritzen et al. 2018). The model runs are completed with the Held–Suarez configuration (Held and Suarez 1994), such that friction exists at the surface and the temperature is relaxed to a prescribed hemispherically symmetric temperature field. The relaxation temperature field is set to equinoctial conditions and there is no absorption of solar energy by the atmosphere (i.e., there are no seasons or diurnal cycles). All runs are performed at T42 resolution with

30 vertical levels, 64 latitude bands, and 128 longitude bands. The simulation is run in the above configuration for one million 6-hourly time steps. The first 20 000 time steps (13.7 yr) are thrown out to account for model spinup.

All data processing is performed to create efficient training data for a CNN (section 2b) to predict the zonally averaged jet response to a range of tropospheric temperature perturbations. Two variables are used in this study, zonally averaged temperature and zonally averaged zonal wind speed. The zonally averaged temperature data are used to calculate the temperature tendency field used as input to the CNN. The zonally averaged zonal wind speed is used to calculate the initial location of the jet and the subsequent shift of the jet, which are used as a CNN input and the CNN prediction, respectively. We exclude data from 200 hPa and above, effectively removing the stratosphere, which is not well resolved in this model without modification (Polvani 2002) and so focus can remain solely on the troposphere for both the forcing and the jet response. Given that this study focuses on hemispheric jets, we take advantage of the hemispheric symmetry in the dry core and use each hemisphere as a separate independent sample, doubling the amount of available training data to two million. After zonally averaging, removing the stratosphere, and considering each hemisphere as a separate sample, the resulting size of the temperature field is 25 vertical levels by 32 latitude bands.

Backward differencing is used to calculate the temperature tendency, which is then smoothed using a backward running mean of 240 time steps (60 days) to remove higher-frequency variability. Removing the high-frequency variability allows the network to focus on learning the response to a forcing that more closely mimics continuous climate change forcing. This smoothing is also aligned with FDT calculations, in which an integration over long time lags (often up to the decorrelation time scale) is done; for example, see Eq. (3) in Hassanzadeh and Kuang (2016b). Smoothing the data before calculating the temperature tendency did not result in any changes in the CNN skill (not shown).

Following established methods (Woollings et al. 2010; McGraw and Barnes 2016), the jet location is defined as the latitude of the maximum wind speed at a pressure level near the surface. Zonal wind speeds from the 850-hPa level are used here and are first smoothed with a 240-time-step (60 days) backward running mean. Then, a second-order polynomial is fit to the peak of the smoothed 850-hPa zonal wind profile and the jet location is defined as the latitude of the maximum wind speeds.

Now that a smoothed zonal temperature tendency and a jet location are calculated, the data are split into training, validation, and testing data. Splitting is completed by chunking the data into three groups where training data is the first chunk, validation the second, and testing the last.

Last, the jet response to a given temperature tendency is defined by the change in jet latitude from the time of input to 120 time steps later (i.e., the jet shift). A positive jet shift indicates a poleward shift in jet location and a negative jet shift indicates an equatorward shift in jet location relative to the jet's latitude at the time of prediction. The jet shift is calculated within each dataset (training, validation, and testing) by subtracting a 240-time-step backward running mean of jet stream locations from a 240-time-step forward running mean of jet stream locations 120 time steps into the future. This processing results in 359 280 training samples, 199 280 validation samples, and 1 399 558 testing samples. Training the CNN required fewer samples than expected because adding more samples to the training dataset did not improve the CNN skill, explaining why the testing dataset is much larger than the training and validation datasets.

b. *Convolutional neural network*

CNNs are commonly used for image recognition and classification tasks as the convolutional layers can extract spatial features in the input image that help the network to learn the correlations between the inputs and output (Fukushima 1980; Yann et al. 1998; Zeiler and Fergus 2014). While a fully connected feedforward network (e.g., LeCun et al. 2015) has the ability to learn the same features extracted by the convolutional layers within a CNN, it may require a larger network and more training data to do so (Yann et al. 1998; Ingrosso and Goldt 2022). In this study, we utilize a CNN so that the network can efficiently learn the correlation between temperature tendencies and the jet response while also trying to minimize the amount of training data required.

The CNN has two inputs: a smoothed temperature tendency field (K day^{-1}) and an initial jet location (degrees latitude). Including the temperature tendency as an input allows us to investigate the jet response to regional temperature tendencies and including the initial jet location supplies the CNN with essential information about the current state of the jet at the prediction time, an important factor for the jet response to forcing (Gerber et al. 2008; Barnes et al. 2010; Kidston and Gerber 2010; Garfinkel et al. 2013). Before the data are input into the CNN, the smoothed temperature tendency field is multiplied by a factor of 10, and the initial jet location is standardized using the standard deviation and mean jet location from the training data. Scaling and standardizing are done so that both inputs have similar magnitudes (order of 1).

The network consists of four convolutional layers: two average pooling layers, three dense layers, and three dropout layers (Fig. 1). Convolutional and dense layers use the hyperbolic tangent activation function. Data are passed through the network as follows: the scaled temperature tendency goes directly into the first convolutional layer with 32 filters of size 3×3 and a stride of 1 followed by a second convolutional layer with the same attributes. The second convolutional layer is then connected to an average pooling layer with a kernel size of 2×2. These three layers, two convolutional and a single average pooling, are repeated with the same attributes with the exception of containing 64 filters rather than 32 in the convolutional layers. The output from the second average pooling layer is flattened and the standardized initial jet location is concatenated to the end. This layer is then fed into the first dense layer with 500 nodes and then goes through a dropout layer with a dropout rate of 30%. The data pass through a combination of dense layers with 500 nodes followed by dropout layers with a dropout rate of 30% two more times. The

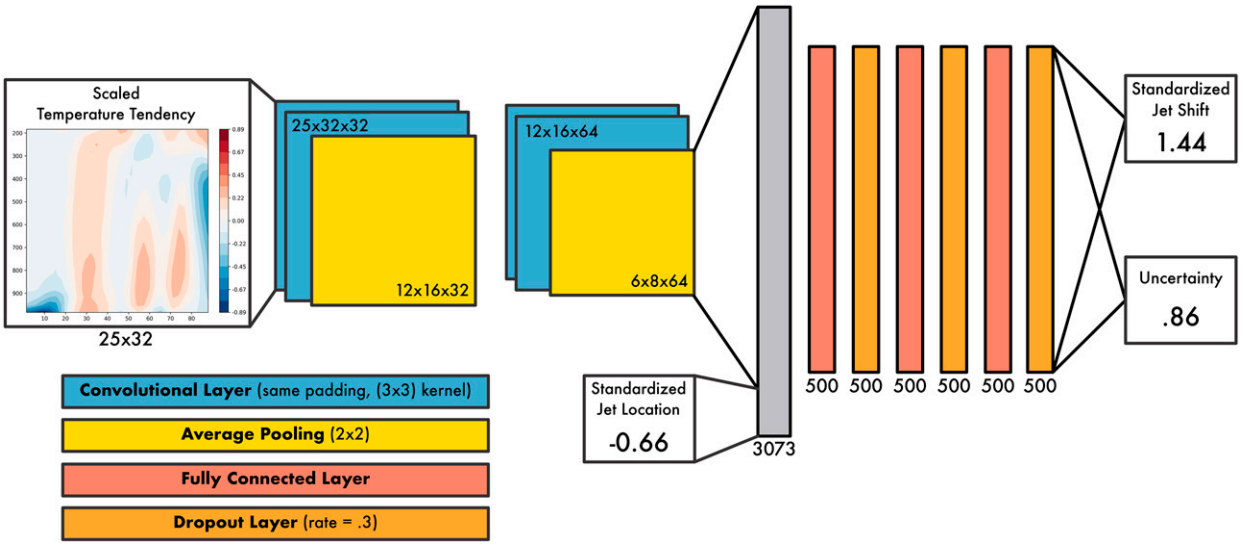

FIG. 1. Schematic of the convolutional neural network with an example of an input and output.

data from the final dropout layer then pass into the output layer consisting of two nodes.

The CNN outputs two values denoted as μ and σ, which represent a mean and standard deviation of a Gaussian distribution where μ denotes the predicted jet shift and σ represents its uncertainty. Estimating values of a distribution is commonly known as maximum likelihood estimation and is regularly used in statistics (Duerr et al. 2020). A neural network that predicts the parameters of a Gaussian distribution is used to quantify network uncertainty (Nix and Weigend 1994a,b), and Gordon and Barnes (2022) recently showed the utility of incorporating uncertainty into a regression neural network for climate science applications. The network learns to predict μ and σ for each sample i through the implementation of the negative log-likelihood loss function:

$$L_i = -\log(p_i), \tag{1}$$

where p is a value of the predicted Gaussian distribution evaluated at the true jet shift for the ith sample. To ensure the network is calibrated we employ the probability integral transform (PIT) probability calibration scheme (Gneiting et al. 2007; Nipen and Stull 2011; Barnes et al. 2023). The PIT histogram for this CNN can be found in the online supplemental material.

To train the CNN, we use the Adam stochastic gradient descent optimization algorithm (Kingma and Ba 2014) with a learning rate of 10^{-7}, a batch size of 256, and a random seed of 300. We apply early stopping to halt the training process once the validation loss fails to decrease for 10 consecutive epochs and restore the model weights to the version with the lowest validation loss (Prechelt 2012).

c. Baselines

We establish two baselines in this study to assess the performance of the CNN and demonstrate that the CNN has learned relationships between the jet response and the regional temperature tendencies. The first baseline is called *persistence*, similar to "persistence forecasting" (MacDonald 1992), where future conditions are predicted to be identical to the current conditions. In our case, this translates to the jet's future location being the same as its location at the time of prediction (i.e., jet shift equal to zero). Comparing this baseline with the CNN ensures that the CNN is predicting jet shifts that are more accurate than predicting a jet shift of zero. The second baseline is called *average evolution* and describes the average movement of the jet based on its position at the time of prediction. For this baseline calculation, the training data is separated into 100 different bins according to the initial jet location, essentially grouping samples with similar initial jet locations together. The average jet shift for each bin is calculated, resulting in an average jet response that is solely dependent on the jet stream's starting position. The average evolution baseline is not sensitive to the number of bins or their exact spacing (not shown). This baseline ensures that the CNN is not just predicting the average evolution of the jet based solely on the initial jet location but is also using the temperature tendency input to make its prediction.

Every test sample is thus associated with three jet shift predictions, one from the CNN and two from the additional persistence and average evolution baselines. Although comparing results between the CNN and the two baselines is useful for placing the CNN's predictions into context, we highlight that the baselines make predictions based solely on information about the initial location of the jet while the CNN is provided additional information in the form of the temperature tendency. Thus, the CNN is able to explore the correlations between a temperature tendency and a jet response.

d. Heating experiments

The main goal of this study is to investigate the jet stream sensitivity to thermal forcing driven by anthropogenic climate

warming. However, as we have designed it, the CNN only trains on data from a long control run (i.e., internal variability), and thus, only provides insights into the forced response if the idea of the FDT holds (Kraichnan 1959; Leith 1975; Marconi et al. 2008). To investigate whether this assumption is valid, we run additional dry core simulations (referred to as *heating experiments*) with zonally symmetric imposed thermal forcing (F) that take the form of a two-dimensional Gaussian in the latitude/pressure plane:

$$F(\Theta, p) = q_o \exp\left[\frac{(|\Theta| - \Theta_o)^2}{\Theta_w^2} - \frac{(|p| - p_o)^2}{p_w^2}\right], \qquad (2)$$

where Θ_o and p_o are the horizontal and vertical centers, respectively; Θ_w and p_w define the width and height; and the magnitude of the forcing is given by q_o. Gaussians that fall near the edge are cut off and therefore are not complete two-dimensional Gaussians. For all 29 Gaussians created in this study, Θ, Θ_w, p, p_w, and q_o are reported in Table 1.

Eighteen heating experiments with either one or two Gaussian thermal forcings imposed in the dry dynamical core are run out to equilibrium to quantify the true jet shift (see experiments 1–18 in Table 1). To perform a direct comparison between the CNN-predicted jet responses and the dry core jet responses, the CNN is given the same temperature tendency that is imposed in each of the dry core heating experiments. For the CNN's initial jet location input, the average jet location from the long control run is used (42.4°). By comparing the true forced response from the dry core with the predicted forced response by the CNN, we are able to investigate the CNN's ability to predict the jet response to thermal forcing from training on internal variability alone. Only experiments 1–18 in Table 1 have true jet responses calculated from forced dry core simulations. All other experiments are used only to evaluate the CNN, and the true jet responses are unknown.

Each heating experiment is initiated at the end of the long control run (time step one million; 684.9 years), and therefore, have the same initial conditions. The heating experiments are run for an additional 20 000 time steps with the first 4000 removed to ensure that the model has reached its new equilibrium. The 850-hPa zonal winds are then used to compute the location of the jet (see section 3a). Last, the true response of the jet to an imposed thermal forcing is defined as the average jet location during the long control run subtracted from the jet location in the corresponding heating experiment.

TABLE 1. Parameters for two-dimensional Gaussians for the forced dry core heating experiments and CNN thermal forcings; "V" indicates a varying parameter. Jet responses to experiments 1–18 have been simulated in the dry dynamical core.

Expt	Θ_0 (° poleward)	Θ_w (° poleward)	p_0 (hPa)	p_w (hPa)	q_0 (K)
1	90	16	1000	250	1.0
	0	27	300	125	−0.1
2	90	16	1000	250	0.5
	0	27	300	125	−0.1
3	0	27	300	125	−0.1
4	90	16	1000	250	−0.5
	0	27	300	125	−0.1
5	90	16	1000	250	−1.0
	0	27	300	125	−0.1
6	90	16	1000	250	1.0
7	90	16	1000	250	0.5
8	90	16	1000	250	−0.5
9	90	16	1000	250	−1.0
10	90	16	1000	250	1.0
	0	27	300	125	0.1
11	90	16	1000	250	0.5
	0	27	300	125	0.1
12	0	27	300	125	0.1
13	90	16	1000	250	−0.5
	0	27	300	125	0.1
14	90	16	1000	250	−1.0
	0	27	300	125	0.1
15	60	15	1000	200	0.25
16	60	15	700	200	0.25
17	30	15	1000	200	0.25
18	30	15	700	200	0.25
19	V	10	V	150	1.0
20	12	15	850	200	V
21	75	15	850	200	V
22	45	15	550	200	V
23	12	15	300	200	V
24	75	15	300	200	V
25	90	16	1000	250	V
	0	27	300	125	V
26	0	27	300	125	V
27	0	27	300	75	V
28	0	13.5	300	125	V
29	0	27	500	125	V

3. Results

a. Evaluation of CNN skill

We begin our discussion of the results with a focus on the deterministic predictions by the CNN (μ). The deterministic skill on the testing data, which we define as the mean absolute error between the predicted jet shift and the true jet shift, reveals how well the CNN generalizes to unseen samples within the control simulation. The first look at the entire testing dataset will appear to show a modest difference; a closer look within the testing dataset will prove more interesting. Figure 2a shows the relationship between the predicted jet shift and the true jet shift where predictions with higher accuracy are closer to the gray diagonal line (one-to-one line). Using orthogonal distance regression (Boggs and Rogers 1990; Virtanen et al. 2020), which takes into account error in both the x and y variables as well as the CNN-predicted uncertainties in y, we calculate the slope from the testing data to be 0.5° poleward/(° poleward). This positive slope demonstrates the CNN has learned relationships between the jet shift and the inputs. However, the slope of the CNN predictions is less than that of the one-to-one line implying that the CNN underestimates the magnitude of the largest jet shifts. This is likely a result of the imbalanced training data as they include more samples with smaller jet shifts than larger ones (shown in Fig. 2a by the density contours). During training, the goal of the CNN is to minimize the negative log-likelihood

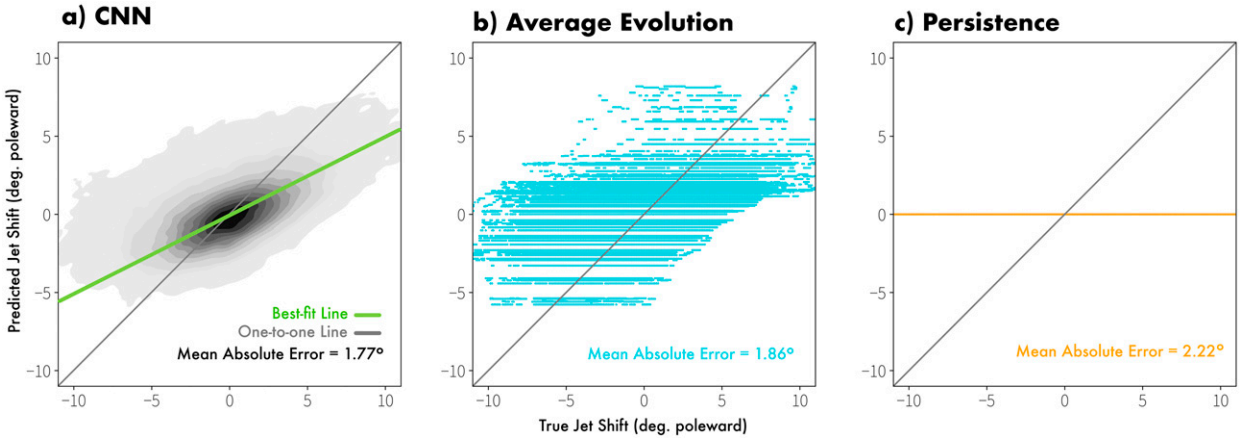

FIG. 2. Predicted jet response (y axis) vs the true jet response (x axis) for the (a) CNN, (b) average evolution baseline, and (c) persistence baseline using the testing data from the control simulation; (a) is a contoured-by-density plot, and (b),(c) scatterplots. Mean absolute errors are shown in the bottom-right corner of each panel. Gray lines represent a perfect prediction (one-to-one line). The green line in (a) represents the best-fit line from the CNN predictions.

loss function [see Eq. (1)], but with an unbalanced dataset, the CNN may never predict the most extreme cases. Applying methods to make the network predict extremes, such as balancing the dataset, using samples weights, or creating custom loss functions (He and Ma 2013; Krawczyk 2016), either caused a severe decrease in skill or did not succeed in solving the problem (not shown). Nonetheless, as we will show next, the CNN outperforms the two benchmark baselines and is an effective tool for exploring jet sensitivity to external forcing.

Comparing the CNN's skill on the testing data with that of the two baselines [average evolution (Fig. 2b) and persistence (Fig. 2c)] allows us to place the CNN's skill into context against other basic prediction methods. Persistence has the lowest performance with a mean absolute error of 2.22°. Average evolution performs only slightly worse than the CNN with mean absolute errors of 1.86° and 1.77°, respectively. Unlike persistence, which can only ever predict a jet shift of zero, average evolution makes a prediction based on the average relationship between the initial jet location and the jet shift of the training data, allowing it to capture the mean jet response. Regardless, average evolution is limited to predicting 1 of the 100 jet shifts resulting from the methods used to calculate it (see section 2): hence the stripes in Fig. 2b. Based on the mean absolute error alone, the CNN outperforms both the persistence and average evolution baselines for the testing data from dry dynamical core long control run.

The mean absolute errors in Fig. 2 represent the error over the entire testing set, which is prone to obscuring interesting details hiding within the distribution. For a more comprehensive analysis of the CNN's skill, the testing data are thus separated into groups based on the initial jet location. The mean absolute error for each group is shown for the CNN and the baselines in Fig. 3 and describes how the CNN's skill and the baselines' skill depend on the initial state of the jet. The gray violin plots behind each bar indicate the CNN's mean absolute error distribution within that bin (i.e., the data used to calculate the CNN mean absolute error in each group). The violin plots are smoothed with a kernel estimator using Scott's rule and 100 points, which are the default parameters of the Matplotlib library (Hunter 2007). The numbers at the bottom of each bar denote the number of samples in that bin. For all bins in Fig. 3, the CNN outperforms the baselines as demonstrated by the CNN's error (black line) falling below the baseline errors (cyan and orange lines). When the initial jet location is equatorward of 42° (labeled as "Avg." along the x axis of Fig. 3), the CNN does considerably better than the baselines, but when the jet location is poleward of 42°, the CNN and average evolution achieve similar skill. That is, in the cases where the initial jet is near the pole, it appears that the CNN does not learn more than average evolution but instead learns this average behavior to make its prediction.

When the initial jet location is near this climatological average position (42.4°), the errors of the CNN and baselines converge (Fig. 3). About 30% of the samples in the training data have initial jet locations within 2° of the climatological average and 14% of these samples have a jet shift between −0.5° and 0.5°. Since so many samples near the climatological average have small jet shifts and because the persistence baseline can only predict a jet shift of zero, the mean absolute error for persistence is at its lowest near the jet's climatological average position. The average evolution baseline converges to a near zero prediction when the initial jet location is near the climatological average, resulting in persistence and average evolution exhibiting similar errors. Although the persistence and average evolution baselines have an advantage near the climatological average, the CNN still outperforms both baselines, implying that the CNN is using the additional information provided by the input temperature tendencies.

Next, we focus on evaluating CNN's ability to predict a jet stream forced response from an artificially constructed idealized temperature tendency not encountered within its noisy training environment. These temperature tendency inputs contain a two-dimensional Gaussian (see Methods) with a prescribed magnitude, size, and location (latitude and pressure).

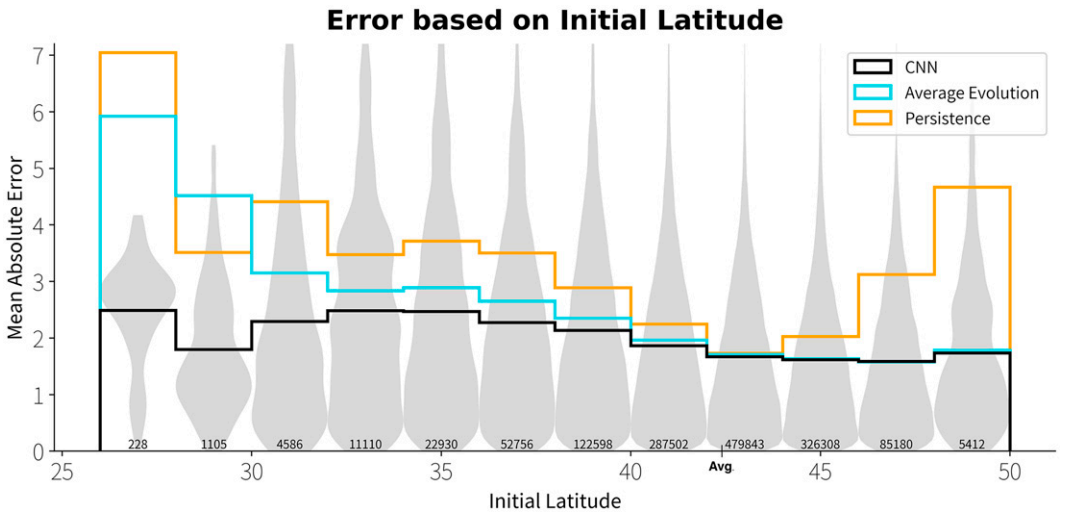

FIG. 3. The mean absolute error from the three prediction methods: CNN (black line), average evolution (cyan line), and persistence (orange line), grouped by initial jet locations. Gray violin plots show the density curves of the CNN's error distribution, where the width corresponds to the frequency of the data. Numbers at the bottom of each bar indicate the number of samples in each group, and the average initial jet location from the training data is marked on the x axis ("Avg"; 42.4°).

Outside of the Gaussian, the temperature tendency field is filled with zeros. Although some of these thermal forcings have magnitudes larger than any temperature tendencies found in the internal variability training data, we will provide strong evidence to support the CNN's ability to extrapolate in the coming sections. CNN predictions made from a thermal forcing use an initial jet location defined by the average jet location of the training data (42.4°). Therefore, differences in predicted jet shifts between temperature tendency inputs are a response to the thermal forcing alone and not the presumed initial state. Exploring sensitivities of the jet response to initial jet location can also be completed. Here we focus on the jet response to a thermal forcing alone.

The shading in Fig. 4 shows the CNN-learned jet sensitivity to the location of heating by holding the magnitude and the shape of a thermal forcing constant and changing only its location (Fig. 4; see experiment 19 in Table 1). An example of a thermal forcing is shown in the gray contours, where the "x"

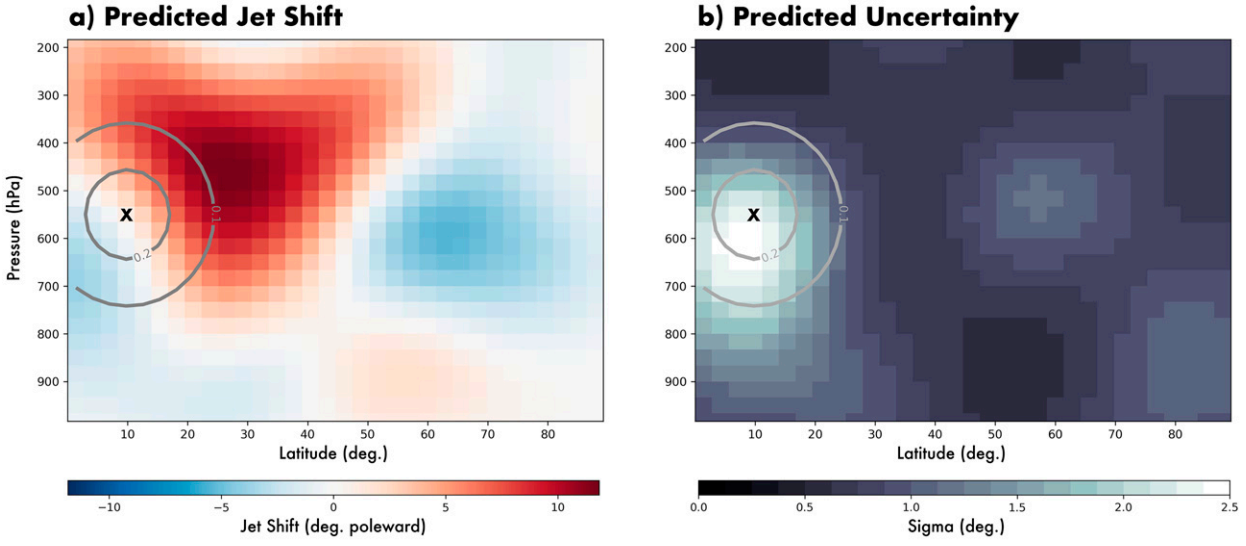

FIG. 4. The thermal forcing with a magnitude of $q_o = 0.25$ K day $^{-1}$ is moved around the latitude and pressure plane where the shading represents the CNN-predicted (a) jet shift and (b) uncertainty, respectively. An example of a thermal forcing is seen in the gray contours in both panels. The "x" marks the center as well as the predicted jet shift and predicted uncertainty associated with that thermal forcing. Gaussian parameters are found in Table 1 (experiment 19).

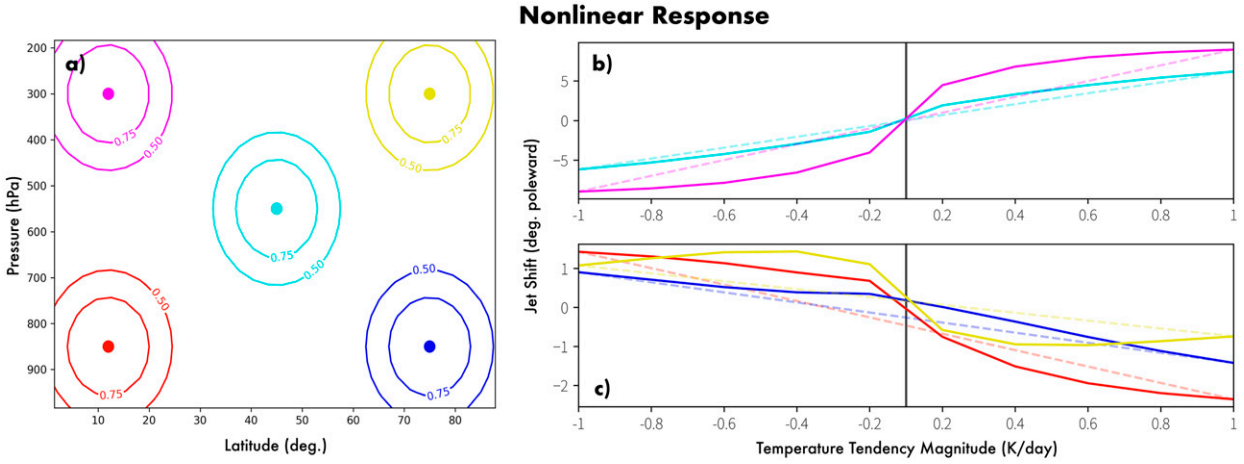

FIG. 5. The CNN learned nonlinear responses: (a) an example thermal forcing with magnitudes of 1 K day^{-1} at the five locations, and (b),(c) the linear response (dashed line) and the predicted response from the CNN (solid lines). Note the different y axes in (b) and (c). Gaussian parameters are found in Table 1 (experiments 20–24).

denotes its center and the color of the shading beneath represents the predicted jet shift (Fig. 4a) and the predicted uncertainty (Fig. 4b) from the temperature tendency.

Figure 4a exhibits multiple known features of the jet response to tropospheric thermal forcings. For example, thermal forcings located higher in the troposphere are known to be more effective at perturbing the jet than thermal forcings located lower in the troposphere (Hassanzadeh and Kuang 2016a; Kim et al. 2021). This feature is learned by the CNN and is shown in Fig. 4a as darker shading at higher pressure levels. In addition, warming in the tropical upper troposphere has been previously shown to cause the jet to shift poleward (Chen et al. 2008; Lim and Simmonds 2009; Butler et al. 2010). This poleward jet shift is seen in Fig. 4a as denoted by the red shading in the tropical upper troposphere. Last, heating at the polar surface has been shown to cause the jet to shift equatorward (Butler et al. 2010; Deser et al. 2010; Screen et al. 2013). This is not seen in Fig. 4a but is seen in later figures (Figs. 5c, 6a,c, and 7). The absence of this feature in Fig. 4a is likely caused by the size of the two-dimensional Gaussian.

Figure 4a also highlights how moving the center of the heating by a few degrees or pressure levels can change the direction of the jet shift. Take for instance heating at the polar surface, where moving the heating from 80° latitude to 75° latitude changes the jet response from an equatorward shift to a poleward shift. Baker et al. (2017) investigates the jet sensitivity to the location of heating by running 306 dry core experiments with an imposed Gaussian shaped temperature tendency that is moved around the latitude-pressure plane, just as we have done here with a trained CNN. In Baker et al. (2017), they show that changes in the latitude of the heating most strongly impact the sign of the jet shift while changing the pressure level has very little impact. Similar behavior is found here with the CNN, although with a few exceptions. A more in-depth discussion about the failure of the CNN to capture the correct direction of the jet shift at the surface of the midlatitudes is found in the online supplemental material.

Recall that the CNN predicts both the jet shift μ as well as its uncertainty σ. Figure 4b displays the predicted uncertainty values and highlights three regions where the CNN is less certain. The CNN is less certain when heating occurs around 10° latitude and 600 hPa ($\sigma \approx 4°$) and additionally has large uncertainty when the heating is centered near 85° latitude and 900 hPa and the 60° latitude and 550 hPa ($\sigma \approx 1.5°$). The reasons behind the greater uncertainty in these regions require further investigation.

b. *Nonlinearities learned by the CNN*

Ideally, a benefit of using a CNN is that it learns a nonlinear relationship between the temperature tendency input and the jet shift output. The hyperbolic tangent activation functions in the convolutional and dense layers of the CNN allow it to learn nonlinear relationships between the inputs and outputs if nonlinearity is present in the data. However, this does not necessarily mean the CNN has learned nonlinear relationships. To evaluate the nonlinearity learned by the CNN we complete two analyses. The first analysis examines how the CNN-predicted jet shift varies as a function of the thermal forcing magnitude. The second analysis looks at scenarios where two thermal forcings are simultaneously present in the temperature tendency input and explores whether the CNN has learned a nonlinear interaction between the two. Keep in mind that neither of these analyses has a ground truth, and so we are exclusively exploring what the CNN has learned. In the next section, we will then further test the accuracy of the CNN with additional dry core simulations.

The first nonlinear analysis explores how the jet shift varies as a function of the thermal forcing magnitude by separately inputting thermal forcings of different magnitudes in five different locations (Fig. 5a; see experiments 20–24 in Table 1). We use 10 different magnitudes that vary from −1.0 to 1.0 K day^{-1} in increments of 0.2 K day^{-1} for each location. For all cases, the initial jet location input is fixed at the average jet location of the training data (42.4°). Figures 5b and 5c compare a linear

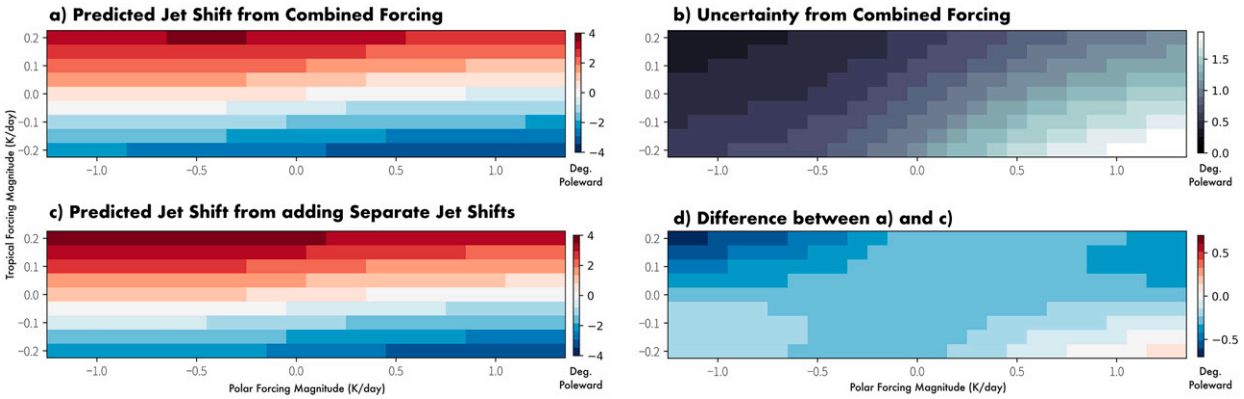

FIG. 6. (a) The CNN-predicted jet shift from when two thermal forcings vary with magnitude, one at the surface of the pole (x axis) and the other in the upper troposphere of the tropics (y axis). (b) The predicted uncertainty for the predictions of (a). (c) As in (a), but the CNN predicts the jet shift from the two thermal forcings independently. (d) The difference between (a) and (c), representing the CNN learned nonlinearity.

relationship (dashed lines; linear response between end points) and the CNN's learned relationship (solid lines) between the thermal forcing magnitude and the jet shift. The jet response to temperature tendencies near the polar surface and the midtropospheric midlatitudes are the most linear as shown by the cyan line in Fig. 5b and the blue line in Fig. 5c. In both of these cases, the CNN-predicted jet shift is most similar to the linear dashed line. In contrast, temperature tendencies in the tropics and the upper troposphere of the polar region have the largest nonlinear response (pink line in Fig. 5b and red line in Fig. 5c; yellow line in Fig. 5c) because these cases vary greatly from the dashed linear line.

We next explore the nonlinearities learned by the CNN when two thermal forcings are present. The thermal forcings are centered on two key regions, the tropical upper troposphere and the polar surface. As discussed previously, warming in the tropical upper troposphere forces the jet to shift poleward (e.g., Chen et al. 2008; Lim and Simmonds 2009; Butler et al. 2010) and warming at the polar surface forces the jet to shift equatorward (e.g., Butler et al. 2010; Deser et al. 2010; Screen et al. 2013). When they occur simultaneously, they force competing effects that can result in a tug-of-war scenario on the jet stream (e.g., Harvey et al. 2015; Chen et al. 2020). To explore the jet sensitivities to this climate change induced tug-of-war, the temperature tendency inputs are composed of a Gaussian thermal forcing at the polar surface and another in the upper troposphere of the tropics. Both vary independently in magnitude during the analyses (see experiment 25 in Table 1); Fig. 6a shows the predicted jet shift μ, and Fig. 6b shows the predicted uncertainty σ for each forcing pattern. With regard to the tug-of-war, studies use a variety of atmospheric models to show that despite opposite forced jet responses, the jet will likely shift poleward (Yin 2005; Harvey et al. 2015). The upper-right quadrant of Fig. 6a depicts the situation in which both thermal forcings are positive (warming). In this scenario, the CNN predicts a poleward shift of the jet in agreement with past work, however, the CNN is not equally certain for all predictions. As shown in Fig. 6b, the CNN is more confident with cooling at the pole and warming in the tropics and less confident with warming in the pole combined with cooling in the tropics. Understanding why these scenarios are more uncertain requires further investigation.

Figure 6a shows the CNN-predicted jet response when two thermal forcings are present in the temperature tendency input. To test whether the CNN has learned a nonlinear impact on the jet from two simultaneous forcings, we task the CNN to predict the jet shift from the two thermal forcings independently (upper tropical troposphere and polar surface) and add the two predicted jet shifts together subsequently. If the CNN exclusively learned a linear response between two forcings, Figs. 6a and 6c would be identical, as predicting a jet shift from combined forcings would be equal to predicting the jet shifts from individual forcing and adding the predictions together. Instead, Fig. 6d shows the difference in predicted jet shifts from these methods and provides evidence of the nonlinearity learned by the CNN where inputs that contain stronger thermal forcings (scenarios in the corners of Fig. 6d) have greater learned nonlinearity.

c. *Out-of-sample tests*

Thus far we have compared the CNN-predicted jet shifts with our established baselines, true jet shifts harvested from the internal variability of the control run, and past work. We next evaluate the ability of the CNN to predict the explicit simulated jet response to an imposed idealized steady thermal forcing outside the training set. The FDT states that the linear response of a nonlinear system to external forcing can be related to the internal variability of the system. Under the assumption that FDT holds, our CNN trained on internal variability may also be able to predict a forced response. We next explore this by comparing the true forced jet shift calculated from additional dry core experiments with the predicted jet shift by the CNN.

We perform 14 additional forced heating experiments with the dry dynamical core (see the methods section and Table 1 experiments 1–14). These 14 heating experiments are motivated

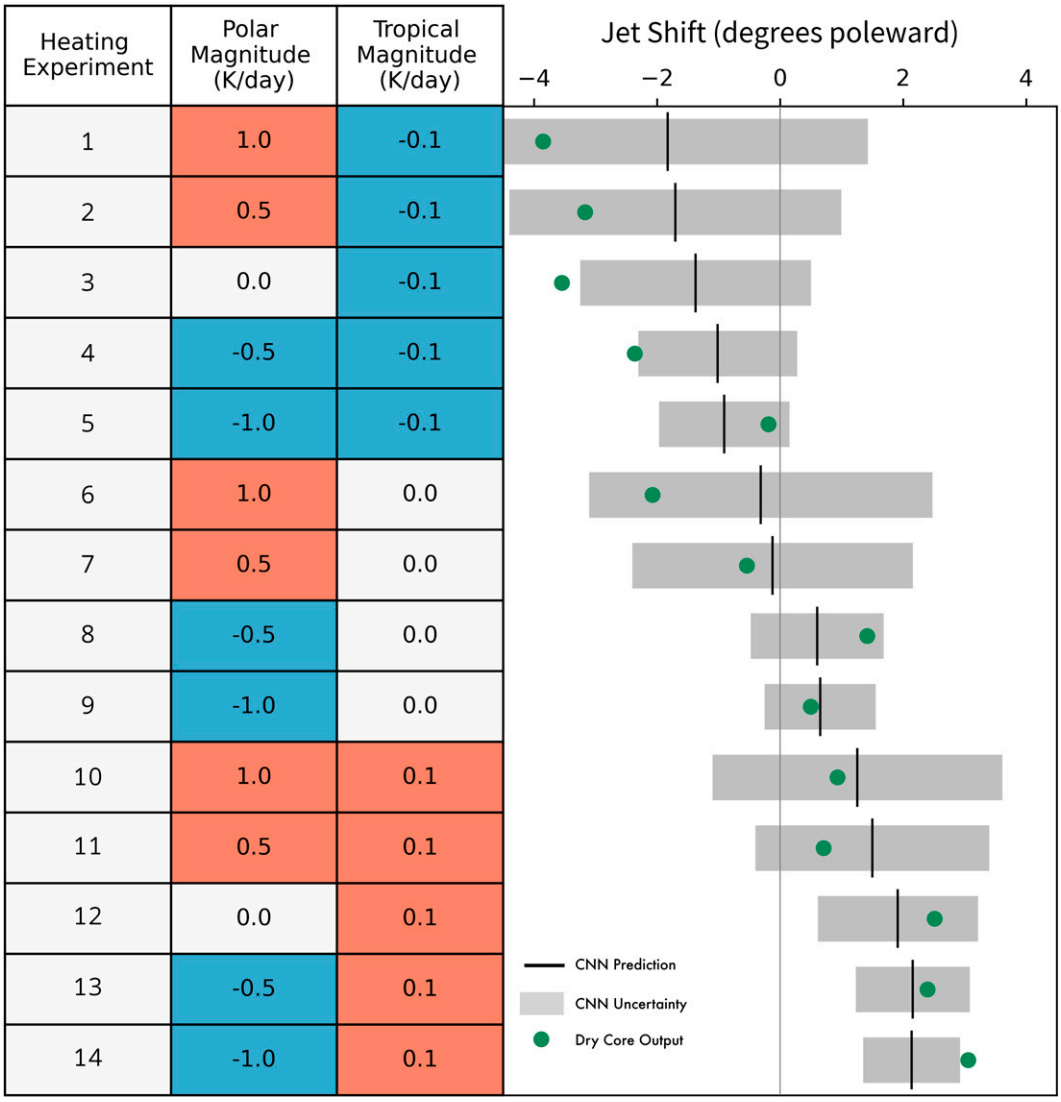

FIG. 7. The 14 heating experiments (experiments 1–14 in Table 1), their resulting jet shift in the dry core, and the jet shift as predicted by the CNN. The left side of the figure includes the magnitude of each Gaussian-shaped thermal forcing. The right side of the figure shows the true forced dry core jet shift (green dots), the predicted CNN jet shift (black line), and the CNN-predicted uncertainty (gray boxes; $\pm 2\sigma$).

by the tug-of-war on the jet resulting from anthropogenic climate change (Harvey et al. 2015; Chen et al. 2020; Stendel et al. 2021). To mimic the tug-of-war, each experiment contains a thermal forcing in the tropical upper troposphere and at the polar surface. The left side of Fig. 7 includes the magnitude of each Gaussian shaped thermal forcing. The forced jet response from the dry dynamical core experiments and the CNN-predicted jet response from 14 heating experiments are shown on the right side of Fig. 7. In comparing the true forced jet shift simulated by the dry core (green dots) and the predicted jet shift by the CNN trained on internal variability (black lines), we see that across all experiments, the CNN accurately captures the sign of the jet shift. Experiments 1–7 exhibit a negative jet shift and 8–14 exhibit a positive jet shift. Furthermore, nearly all of the experiments (excluding 3, 4, and 14) have forced jet shifts that fall well within the uncertainty bounds predicted by the CNN ($\pm 2\sigma$; gray boxes).

Heating experiments 6, 7, 8, and 9 contain only a thermal forcing at the polar surface (no thermal forcing in the tropical upper troposphere) and are therefore useful for investigating the difference in CNN-predicted uncertainty between polar warming and polar cooling. In heating experiments 6 and 7, which contain polar warming, the CNN is less certain (larger σ), in contrast to heating experiments 8 and 9, which contains polar cooling, where the CNN is more certain (smaller σ).

The CNN's uncertainty when there is a thermal forcing in the upper troposphere of the tropics is more difficult to discern from Fig. 7 because the CNN's uncertainty is impacted

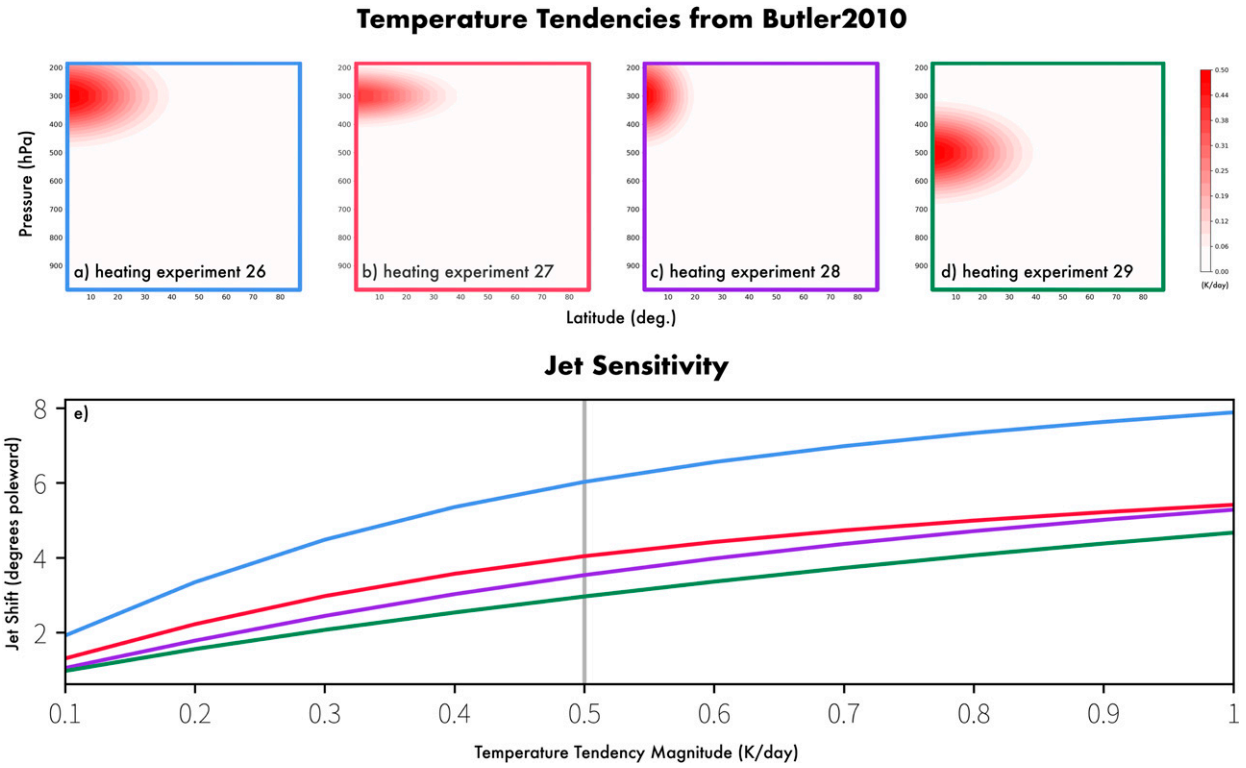

FIG. 8. (a)–(d) Example thermal forcings of magnitude 0.5 K day^{-1} respectively representing the four heating experiments used in B10; Gaussian parameters are found in Table 1 (experiments 26–29). The colors of the panel outlines correspond to the predicted jet shifts (y axis) in (e), which are shown as a function of the magnitude of the temperate forcing (x axis). The vertical gray line at 0.5 K day^{-1} corresponds to the magnitude of the heating experiments used in B10.

considerably by the thermal forcing at the polar surface. However, heating experiments 3 and 12 include only a thermal forcing in the tropical upper troposphere, one cooling and one warming, respectively. These heating experiments suggest that the CNN is less certain when the upper tropical troposphere is cooling rather than warming.

4. Implications for sensitivity analysis

Given the CNN's ability to replicate the sign of the jet stream's forced response as validated with the additional forced dry dynamical core experiments, we propose that our approach can be deployed as a computationally efficient tool to aid in the design of forced model experiments. To demonstrate this, we next revisit a historical study (Butler et al. 2010, hereinafter B10) and show how the pretrained CNN can be used to replicate the study's results as well as to document possible sensitivities not included in the initial work. In B10, differences in dry core atmospheric circulations due to variations in the location and shape of thermal forcings were identified and documented. To test the atmospheric sensitivity, B10 ran multiple experiments with different imposed thermal forcing patterns in the Colorado State University general circulation model (Ringler et al. 2000). Although B10 did not quantify a shift of the jet stream the same way as we do here, the study showed the zonal-mean zonal wind response and discussed which heating experiments resulted in a stronger wind response. From this information, we are able to infer the relative magnitude of the jet shift for each experiment.

Here, we focus on four specific heating experiments within B10 that aim to investigate how sensitive the circulation response is to the height and shape of tropical upper-tropospheric heating. Examples of the thermal forcing patterns imposed in B10 are shown in Figs. 8a–d and will be referred to as heating experiments 26, 27, 28, and 29 for this discussion. In the original study, B10 found that the jet shifts poleward in response to all heating experiments, but that the magnitudes of the shifts varied. B10 shows that heating experiment 26 had the strongest jet response and compressing the forcing vertically (heating experiment 27) or compressing the forcing in the meridional direction (heating experiment 28) weakened the wind response. In the last experiment (heating experiment 29), B10 showed that when the forcing was compressed vertically and moved lower in the troposphere, the wind response was even weaker.

Using values reported in Table 1 of B10, we recreated the thermal forcing patterns (see experiments 26–29 in Table 1) and input them into the CNN with the initial jet location set to the average jet location of the training data (42.4°). In addition to the magnitude of heating used in B10 (q_o = 0.5 K day^{-1}), here the jet sensitivity to the magnitude of the thermal forcing is also included since it is trivial to explore once the CNN is trained. Figure 8e shows the predicted jet shift from the four heating

experiments with varying magnitudes, and the vertical gray line indicates the 0.5 K day $^{-1}$ magnitude used in B10. The CNN predicts the same relative relationship between the heating experiments as found in B10, where heating experiment 26 exhibits the strongest jet response and heating experiment 29 exhibits the weakest. Furthermore, by exploring the jet sensitivity to the magnitude of heating, Fig. 8e shows new information about jet sensitivity. For example, as the magnitude of the thermal forcing increases, heating experiment 27 (compressed vertically) and heating experiment 28 (compressed meridionally) converge to the same jet shift. Alternatively, when the magnitude of the thermal forcing decreases, heating experiment 28 and heating experiment 29 (compressed vertically and lower in the troposphere) converge to the same jet shift. These two results could be confirmed with a few targeted forced dry core simulations (not done here), though it was the ability provided by the CNN to quickly explore jet shifts in response to thermal forcings that allowed us to discover these possible jet sensitivities.

In this section, we show a successful example of using the CNN to explore the jet sensitivities inside the dry dynamical core. The CNN is not perfect in its predictions, which may be due to a lack of predictability, nonoptimal training of the CNN, a breakdown of the FDT, or a combination of the three. However, as demonstrated throughout this paper, the comparisons in the CNN-predicted jet shifts have the same sign and similar magnitudes to the forced jet shifts from the dry core in response to a range of temperature tendencies and once trained, can make predictions quickly. We emphasize that this method should not replace the need to run designed climate model experiments. Rather, training CNN on an existing long control run could provide the opportunity to explore a large number of forcing experiments before any forced model runs are simulated, and it could be especially helpful for planning forced experiments in dynamic model simulations.

5. Conclusions

We explore the jet stream's response to external forcing by training a CNN on smoothed temperature tendencies from a dry dynamical core long control run to predict a shift in the jet stream's location 30 days later. The main motivation of this work is to explore the potential for training a CNN on internal variability alone and then using it to examine possible nonlinear responses of the jet to tropospheric thermal forcing that more closely resemble anthropogenic climate change. Because the CNN is trained entirely on data from a control simulation, it exclusively learns from internal variability. Nevertheless, by comparing the CNN-predicted jet shifts with established baselines, peer reviewed literature, and additional dry core heating experiments, we show that the CNN can predict the forced jet shift to sustained forcing. The trained CNN is then used to investigate jet sensitivities to scenarios that mimic the tug-of-war between the tropics and poles under anthropogenic climate change. Given the CNN's ability to predict the jet response to thermal forcings, we propose training a CNN on long control runs that are increasingly becoming more available to explore model sensitivities to various forcings as a tool to aid in early-stage climate model experiment design. Future work could include extending this method to evaluate whether it generalizes to different experimental frameworks, including, but not limited to, evaluating it with three-dimensional predictors, training on coupled climate model simulations, and learning complex nonlinear climate responses from forced simulations where FDT does not hold.

Acknowledgments. Authors Connolly and Barnes are supported, in part, by NSF CAREER AGS-1749261 under the Climate and Large-Scale Dynamics program. The authors thank Marybeth Arcodia for providing valuable support during the writing process.

Data availability statement. The code used to prepare the data can be found on Zenodo (https://doi.org/10.5281/zenodo.7796266). Data are also available on Zenodo (https://doi.org/10.5281/zenodo.7796170).

REFERENCES

Abbot, D. S., and E. Tziperman, 2008: A high-latitude convective cloud feedback and equable climates. *Quart. J. Roy. Meteor. Soc.*, **134**, 165–185, https://doi.org/10.1002/qj.211.

Athanasiadis, P. J., J. M. Wallace, and J. J. Wettstein, 2010: Patterns of wintertime jet stream variability and their relation to the storm tracks. *J. Atmos. Sci.*, **67**, 1361–1381, https://doi.org/10.1175/2009JAS3270.1.

Baker, H. S., T. Woollings, and C. Mbengue, 2017: Eddy-driven jet sensitivity to diabatic heating in an idealized GCM. *J. Climate*, **30**, 6413–6431, https://doi.org/10.1175/JCLI-D-16-0864.1.

Barnes, E. A., and L. Polvani, 2013: Response of the midlatitude jets, and of their variability, to increased greenhouse gases in the CMIP5 models. *J. Climate*, **26**, 7117–7135, https://doi.org/10.1175/JCLI-D-12-00536.1.

——, D. L. Hartmann, D. M. W. Frierson, and J. Kidston, 2010: Effect of latitude on the persistence of eddy-driven jets. *Geophys. Res. Lett.*, **37**, L11804, https://doi.org/10.1029/2010GL043199.

——, R. J. Barnes, and M. DeMaria, 2023: Sinh-arcsinh-normal distributions to add uncertainty to neural network regression tasks: Applications to tropical cyclone intensity forecasts. *Environ. Data Sci.*, in press.

Bibi, A., K. Ullah, Z. Yushu, Z. Wang, and S. Gao, 2020: Role of westerly jet in torrential rainfall during monsoon over northern Pakistan. *Earth Space Sci.*, **7**, e2019EA001022, https://doi.org/10.1029/2019EA001022.

Blackport, R., and J. A. Screen, 2020: Insignificant effect of Arctic amplification on the amplitude of midlatitude atmospheric waves. *Sci. Adv.*, **6**, eaay2880, https://doi.org/10.1126/sciadv.aay2880.

Blunden, J., and D. S. Arndt, 2012: State of the Climate in 2011. *Bull. Amer. Meteor. Soc.*, **93** (7), S1–S282, https://doi.org/10.1175/2012BAMSStateoftheClimate.1.

Boggs, P. T., and J. E. Rogers, 1990: Orthogonal distance regression. *Contemp. Math.*, **112**, 183–194, https://doi.org/10.1090/conm/112/1087109.

Butler, A. H., D. W. J. Thompson, and R. Heikes, 2010: The steady-state atmospheric circulation response to climate change–like thermal forcings in a simple general circulation model. *J. Climate*, **23**, 3474–3496, https://doi.org/10.1175/2010JCLI3228.1.

Chen, G., J. Lu, and D. M. W. Frierson, 2008: Phase speed spectra and the latitude of surface westerlies: Interannual variability and global warming trend. *J. Climate*, **21**, 5942–5959, https://doi.org/10.1175/2008JCLI2306.1.

——, P. Zhang, and J. Lu, 2020: Sensitivity of the latitude of the westerly jet stream to climate forcing. *Geophys. Res. Lett.*, **47**, e2019GL086563, https://doi.org/10.1029/2019GL086563.

Cohen, J., and Coauthors, 2014: Recent Arctic amplification and extreme mid-latitude weather. *Nat. Geosci.*, **7**, 627–637, https://doi.org/10.1038/ngeo2234.

Cooper, F. C., and P. H. Haynes, 2011: Climate sensitivity via a nonparametric fluctuation–dissipation theorem. *J. Atmos. Sci.*, **68**, 937–953, https://doi.org/10.1175/2010JAS3633.1.

Coumou, D., and S. Rahmstorf, 2012: A decade of weather extremes. *Nat. Climate Change*, **2**, 491–496, https://doi.org/10.1038/nclimate1452.

Dai, A., D. Luo, M. Song, and J. Liu, 2019: Arctic amplification is caused by sea-ice loss under increasing CO_2. *Nat. Commun.*, **10**, 121, https://doi.org/10.1038/s41467-018-07954-9.

Deser, C., R. Tomas, M. Alexander, and D. Lawrence, 2010: The seasonal atmospheric response to projected Arctic sea ice loss in the late twenty-first century. *J. Climate*, **23**, 333–351, https://doi.org/10.1175/2009JCLI3053.1.

Druckenmiller, M. L., and Coauthors, 2021: The Arctic [in "State of the Climate in 2020"]. *Bull. Amer. Meteor. Soc.*, **102** (8), S263–S316, https://doi.org/10.1175/BAMS-D-21-0086.1.

Duerr, O., B. Sick, and E. Murina, 2020: *Probabilistic Deep Learning: With Python, Keras and TensorFlow Probability*. 1st ed. Manning, 296 pp.

Dunn-Sigouin, E., C. Li, and P. J. Kushner, 2021: Limited influence of localized tropical sea-surface temperatures on moisture transport into the Arctic. *Geophys. Res. Lett.*, **48**, e2020GL091540, https://doi.org/10.1029/2020GL091540.

Fuchs, D., S. Sherwood, and D. Hernandez, 2015: An exploration of multivariate fluctuation dissipation operators and their response to sea surface temperature perturbations. *J. Atmos. Sci.*, **72**, 472–486, https://doi.org/10.1175/JAS-D-14-0077.1.

Fukushima, K., 1980: Neocognitron: A self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.*, **36**, 193–202, https://doi.org/10.1007/BF00344251.

Garfinkel, C. I., D. W. Waugh, and E. P. Gerber, 2013: The effect of tropospheric jet latitude on coupling between the stratospheric polar vortex and the troposphere. *J. Climate*, **26**, 2077–2095, https://doi.org/10.1175/JCLI-D-12-00301.1.

Gerber, E. P., S. Voronin, and L. M. Polvani, 2008: Testing the annular mode autocorrelation time scale in simple atmospheric general circulation models. *Mon. Wea. Rev.*, **136**, 1523–1536, https://doi.org/10.1175/2007MWR2211.1.

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.*, **69B**, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Gordon, E. M., and E. A. Barnes, 2022: Incorporating uncertainty into a regression neural network enables identification of decadal state-dependent predictability in CESM2. *Geophys. Res. Lett.*, **49**, e2022GL098635, https://doi.org/10.1029/2022GL098635.

Graversen, R. G., and P. L. Langen, 2019: On the role of the atmospheric energy transport in 2 × CO_2–induced polar amplification in CESM1. *J. Climate*, **32**, 3941–3956, https://doi.org/10.1175/JCLI-D-18-0546.1.

Grise, K. M., and L. M. Polvani, 2016: Is climate sensitivity related to dynamical sensitivity? *J. Geophys. Res. Atmos.*, **121**, 5159–5176, https://doi.org/10.1002/2015JD024687.

Gritsun, A., and G. Branstator, 2007: Climate response using a three-dimensional operator based on the fluctuation–dissipation theorem. *J. Atmos. Sci.*, **64**, 2558–2575, https://doi.org/10.1175/JAS3943.1.

Harvey, B. J., L. C. Shaffrey, and T. J. Woollings, 2015: Deconstructing the climate change response of the Northern Hemisphere wintertime storm tracks. *Climate Dyn.*, **45**, 2847–2860, https://doi.org/10.1007/s00382-015-2510-8.

Hassanzadeh, P., and Z. Kuang, 2016a: The linear response function of an idealized atmosphere. Part I: Construction using Green's functions and applications. *J. Atmos. Sci.*, **73**, 3423–3439, https://doi.org/10.1175/JAS-D-15-0338.1.

——, and ——, 2016b: The linear response function of an idealized atmosphere. Part II: Implications for the practical use of the fluctuation–dissipation theorem and the role of operator's nonnormality. *J. Atmos. Sci.*, **73**, 3441–3452, https://doi.org/10.1175/JAS-D-16-0099.1.

——, ——, and B. F. Farrell, 2014: Responses of midlatitude blocks and wave amplitude to changes in the meridional temperature gradient in an idealized dry GCM. *Geophys. Res. Lett.*, **41**, 5223–5232, https://doi.org/10.1002/2014GL060764.

He, H., and Y. Ma, 2013: *Imbalanced Learning: Foundations, Algorithms, and Applications*. John Wiley and Sons, 340 pp.

Held, I. M., 1993: Large-scale dynamics and global warming. *Bull. Amer. Meteor. Soc.*, **74**, 228–242, https://doi.org/10.1175/1520-0477(1993)074<0228:LSDAGW>2.0.CO;2.

——, and M. J. Suarez, 1994: A proposal for the intercomparison of the dynamical cores of atmospheric general circulation models. *Bull. Amer. Meteor. Soc.*, **75**, 1825–1830, https://doi.org/10.1175/1520-0477(1994)075<1825:APFTIO>2.0.CO;2.

Hunter, J. D., 2007: Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95, https://doi.org/10.1109/MCSE.2007.55.

Hwang, Y.-T., and D. M. W. Frierson, 2010: Increasing atmospheric poleward energy transport with global warming. *Geophys. Res. Lett.*, **37**, L24807, https://doi.org/10.1029/2010GL045440.

Ingrosso, A., and S. Goldt, 2022: Data-driven emergence of convolutional structure in neural networks. *Proc. Natl. Acad. Sci. USA*, **119**, e2201854119, https://doi.org/10.1073/pnas.2201854119.

Kattsov, V. M., V. E. Ryabinin, J. E. Overland, M. C. Serreze, M. Visbeck, J. E. Walsh, W. Meier, and X. Zhang, 2010: Arctic sea-ice change: A grand challenge of climate science. *J. Glaciol.*, **56**, 1115–1121, https://doi.org/10.3189/002214311796406176.

Khodkar, M. A., and P. Hassanzadeh, 2018: Data-driven reduced modelling of turbulent Rayleigh–Bénard convection using DMD-enhanced fluctuation–dissipation theorem. *J. Fluid Mech.*, **852**, R3, https://doi.org/10.1017/jfm.2018.586.

Kidston, J., and E. P. Gerber, 2010: Intermodel variability of the poleward shift of the austral jet stream in the CMIP3 integrations linked to biases in 20th century climatology. *Geophys. Res. Lett.*, **37**, L09708, https://doi.org/10.1029/2010GL042873.

Kim, D., S. M. Kang, T. M. Merlis, and Y. Shin, 2021: Atmospheric circulation sensitivity to changes in the vertical structure of polar warming. *Geophys. Res. Lett.*, **48**, e2021GL094726, https://doi.org/10.1029/2021GL094726.

Kingma, D. P., and J. Ba, 2014: Adam: A method for stochastic optimization. arXiv, 1412.6980v9, https://doi.org/10.48550/arXiv.1412.6980.

Kornhuber, K., D. Coumou, E. Vogel, C. Lesk, J. F. Donges, J. Lehmann, and R. M. Horton, 2020: Amplified Rossby waves enhance risk of concurrent heatwaves in major breadbasket regions. *Nat. Climate Change*, **10**, 48–53, https://doi.org/10.1038/s41558-019-0637-z.

Kraichnan, R. H., 1959: Classical fluctuation-relaxation theorem. *Phys. Rev.*, **113**, 1181–1182, https://doi.org/10.1103/PhysRev.113.1181.

Krawczyk, B., 2016: Learning from imbalanced data: Open challenges and future directions. *Prog. Artif. Intell.*, **5**, 221–232, https://doi.org/10.1007/s13748-016-0094-0.

Lauritzen, P. H., and Coauthors, 2018: NCAR release of CAM-SE in CESM2.0: A reformulation of the spectral element dynamical core in dry-mass vertical coordinates with comprehensive treatment of condensates and energy. *J. Adv. Model. Earth Syst.*, **10**, 1537–1570, https://doi.org/10.1029/2017MS001257.

LeCun, Y., Y. Bengio, and G. Hinton, 2015: Deep learning. *Nature*, **521**, 436–444, https://doi.org/10.1038/nature14539.

Lee, S., 2014: A theory for polar amplification from a general circulation perspective. *Asia-Pac. J. Atmos. Sci.*, **50**, 31–43, https://doi.org/10.1007/s13143-014-0024-7.

Leith, C. E., 1975: Climate response and fluctuation dissipation. *J. Atmos. Sci.*, **32**, 2022–2026, https://doi.org/10.1175/1520-0469(1975)032<2022:CRAFD>2.0.CO;2.

Lim, E.-P., and I. Simmonds, 2009: Effect of tropospheric temperature change on the zonal mean circulation and SH winter extratropical cyclones. *Climate Dyn.*, **33**, 19–32, https://doi.org/10.1007/s00382-008-0444-0.

Lutsko, N. J., I. M. Held, and P. Zurita-Gotor, 2015: Applying the fluctuation–dissipation theorem to a two-layer model of quasi-geostrophic turbulence. *J. Atmos. Sci.*, **72**, 3161–3177, https://doi.org/10.1175/JAS-D-14-0356.1.

MacDonald, G., 1992: Persistence in climate. The MITRE Corporation Tech. Rep. JSR-91-340, 70 pp., https://apps.dtic.mil/sti/pdfs/ADA247632.pdf.

Madonna, E., C. Li, C. M. Grams, and T. Woollings, 2017: The link between eddy-driven jet variability and weather regimes in the North Atlantic–European sector. *Quart. J. Roy. Meteor. Soc.*, **143**, 2960–2972, https://doi.org/10.1002/qj.3155.

Mahlstein, I., O. Martius, C. Chevalier, and D. Ginsbourger, 2012: Changes in the odds of extreme events in the Atlantic basin depending on the position of the extratropical jet. *Geophys. Res. Lett.*, **39**, L22805, https://doi.org/10.1029/2012GL053993.

Majda, A., R. V. Abramov, and M. J. Grote, 2005: *Information Theory and Stochastics for Multiscale Nonlinear Systems. CRM Monogr. Series*, Vol. 25, American Mathematical Society, 133 pp.

Manabe, S., and R. J. Stouffer, 1980: Sensitivity of a global climate model to an increase of CO_2 concentration in the atmosphere. *J. Geophys. Res.*, **85**, 5529–5554, https://doi.org/10.1029/JC085iC10p05529.

——, D. G. Hahn, and J. Leith Holloway Jr., 1974: The seasonal variation of the tropical circulation as simulated by a global model of the atmosphere. *J. Atmos. Sci.*, **31**, 43–83, https://doi.org/10.1175/1520-0469(1974)031<0043:TSVOTT>2.0.CO;2.

Marconi, U. M. B., A. Puglisi, L. Rondoni, and A. Vulpiani, 2008: Fluctuation–dissipation: Response theory in statistical physics. *Phys. Rep.*, **461**, 111–195, https://doi.org/10.1016/j.physrep.2008.02.002.

Mattingly, K. S., J. T. McLeod, J. A. Knox, J. M. Shepherd, and T. L. Mote, 2015: A climatological assessment of Greenland blocking conditions associated with the track of Hurricane Sandy and historical North Atlantic Hurricanes. *Int. J. Climatol.*, **35**, 746–760, https://doi.org/10.1002/joc.4018.

Mbengue, C., and T. Schneider, 2013: Storm track shifts under climate change: What can be learned from large-scale dry dynamics. *J. Climate*, **26**, 9923–9930, https://doi.org/10.1175/JCLI-D-13-00404.1.

McGraw, M. C., and E. A. Barnes, 2016: Seasonal sensitivity of the eddy-driven jet to tropospheric heating in an idealized AGCM. *J. Climate*, **29**, 5223–5240, https://doi.org/10.1175/JCLI-D-15-0723.1.

Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor, 2007: The WCRP CMIP3 multimodel dataset: A new era in climate change research. *Bull. Amer. Meteor. Soc.*, **88**, 1383–1394, https://doi.org/10.1175/BAMS-88-9-1383.

Nakamura, H., T. Sampe, Y. Tanimoto, and A. Shimpo, 2004: Observed associations among storm tracks, jet streams and midlatitude oceanic fronts. *Earth's Climate: The Ocean–Atmosphere Interaction*, *Geophys. Monogr.*, Vol. 147, Amer. Geophys. Union, 329–345, https://doi.org/10.1029/147GM18.

Nipen, T., and R. Stull, 2011: Calibrating probabilistic forecasts from an NWP ensemble. *Tellus*, **63A**, 858–875, https://doi.org/10.1111/j.1600-0870.2011.00535.x.

Nix, D. A., and A. S. Weigend, 1994a: Estimating the mean and variance of the target probability distribution. *Proc. 1994 IEEE Int. Conf. on Neural Networks (ICNN'94)*, Vol. 1, Orlando, FL, Institute of Electrical and Electronics Engineers, 55–60, https://doi.org/10.1109/ICNN.1994.374138.

——, and ——, 1994b: Learning local error bars for nonlinear regression. *NIPS'94: Proc. Seventh Int. Conf. on Neural Information Processing Systems*, Denver, CO, Association for Computing Machinery, 489–496, https://dl.acm.org/doi/abs/10.5555/2998687.2998748.

Phipps, S. J., 2010: The CSIRO Mk3L climate system model v1.2. CRC Tech. Rep., 122 pp., https://www.stevenphipps.com/publications/phipps2010b.pdf.

Pithan, F., and T. Mauritsen, 2014: Arctic amplification dominated by temperature feedbacks in contemporary climate models. *Nat. Geosci.*, **7**, 181–184, https://doi.org/10.1038/ngeo2071.

Polvani, L. M., 2002: Tropospheric response to stratospheric perturbations in a relatively simple general circulation model. *Geophys. Res. Lett.*, **29**, 1114, https://doi.org/10.1029/2001GL014284.

Prechelt, L., 2012: Early stopping—but when? *Neural Networks: Tricks of the Trade*, G. Montavon, G. B. Orr, and K.-R. Müller, Eds., Lecture Notes in Computer Science, Vol. 7700, Springer 53–67.

Ringler, T. D., R. P. Heikes, and D. A. Randall, 2000: Modeling the atmospheric general circulation using a spherical geodesic grid: A new class of dynamical cores. *Mon. Wea. Rev.*, **128**, 2471–2490, https://doi.org/10.1175/1520-0493(2000)128<2471:MTAGCU>2.0.CO;2.

Röthlisberger, M., S. Pfahl, and O. Martius, 2016: Regional-scale jet waviness modulates the occurrence of midlatitude weather extremes. *Geophys. Res. Lett.*, **43**, 10 989–10 997, https://doi.org/10.1002/2016GL070944.

Rousi, E., F. Selten, S. Rahmstorf, and D. Coumou, 2021: Changes in North Atlantic atmospheric circulation in a warmer climate favor winter flooding and summer drought over Europe. *J. Climate*, **34**, 2277–2295, https://doi.org/10.1175/JCLI-D-20-0311.1.

——, K. Kornhuber, G. Beobide-Arsuaga, F. Luo, and D. Coumou, 2022: Accelerated western European heatwave trends linked to more-persistent double jets over Eurasia. *Nat. Commun.*, **13**, 3851, https://doi.org/10.1038/s41467-022-31432-y.

Schubert, S., H. Wang, and M. Suarez, 2011: Warm season subseasonal variability and climate extremes in the Northern Hemisphere: The role of stationary Rossby waves. *J. Climate*, **24**, 4773–4792, https://doi.org/10.1175/JCLI-D-10-05035.1.

Screen, J. A., I. Simmonds, C. Deser, and R. Tomas, 2013: The atmospheric response to three decades of observed Arctic sea

ice loss. *J. Climate*, **26**, 1230–1248, https://doi.org/10.1175/JCLI-D-12-00063.1.

Shaw, T. A., and Coauthors, 2016: Storm track processes and the opposing influences of climate change. *Nat. Geosci.*, **9**, 656–664, https://doi.org/10.1038/ngeo2783.

Sherwood, S. C., and N. Nishant, 2015: Atmospheric changes through 2012 as shown by iteratively homogenized radiosonde temperature and wind data (IUKv2). *Environ. Res. Lett.*, **10**, 054007, https://doi.org/10.1088/1748-9326/10/5/054007.

Stendel, M., J. Francis, R. White, P. D. Williams, and T. Woollings, 2021: The jet stream and climate change. *Climate Change*, 3rd ed. T. M. Letcher, Ed., Elsevier, 327–357.

Swart, N. C., and J. C. Fyfe, 2012: Observed and simulated changes in the Southern Hemisphere surface westerly wind-stress. *Geophys. Res. Lett.*, **39**, L16711, https://doi.org/10.1029/2012GL052810.

Tjernström, M., J. Sedlar, and M. D. Shupe, 2008: How well do regional climate models reproduce radiation and clouds in the Arctic? An evaluation of ARCMIP simulations. *J. Appl. Meteor. Climatol.*, **47**, 2405–2422, https://doi.org/10.1175/2008JAMC1845.1.

Vihma, T., 2014: Effects of Arctic sea ice decline on weather and climate: A review. *Surv. Geophys.*, **35**, 1175–1214, https://doi.org/10.1007/s10712-014-9284-0.

Virtanen, P., and Coauthors, 2020: SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nat. Methods*, **17**, 261–272, https://doi.org/10.1038/s41592-019-0686-2.

Woollings, T., A. Hannachi, and B. Hoskins, 2010: Variability of the North Atlantic eddy-driven jet stream. *Quart. J. Roy. Meteor. Soc.*, **136**, 856–868, https://doi.org/10.1002/qj.625.

Yann, L., B. Leon, B. Yoshia, and P. Haffner, 1998: Gradient-based learning applied to document recognition. *Proc. IEEE*, **86**, 2278–2324, https://doi.org/10.1109/5.726791.

Yin, J. H., 2005: A consistent poleward shift of the storm tracks in simulations of 21st century climate. *Geophys. Res. Lett.*, **32**, L18701, https://doi.org/10.1029/2005GL023684.

Zeiler, M. D., and R. Fergus, 2014: Visualizing and understanding convolutional networks. *Computer Vision—ECCV 2014*, D. Fleet et al., Eds., Lecture Notes in Computer Science, Vol. 8689, Springer, 818–833.