Using Explainable AI and Transfer Learning to understand and predict the maintenance of Atlantic blocking with limited observational data

Huan Zhang[1], Justin Finkel[2], Dorian S. Abbot[3], Edwin P. Gerber[1], and Jonathan Weare[1]

[1]Courant Institute of Mathematical Sciences, New York University
[2]Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology
[3]Department of the Geophysical Sciences, University of Chicago

Key Points:

- Given sufficient training data, convolutional neural networks can predict the maintenance of Atlantic blocking from an initial blocked state.
- Transfer learning from an idealized model to reanalysis data enables predictive skill in the low data regime of the observational record.
- Feature importance analysis reveals the influence of upstream flow on blocking persistence and quantifies biases in the idealized model.

Corresponding author: Jonathan Weare, `weare@nyu.edu`

Abstract

Blocking events are an important cause of extreme weather, especially long-lasting blocking events that trap weather systems in place. The duration of blocking events is, however, underestimated in climate models. Explainable Artificial Intelligence are a class of data analysis methods that can help identify physical causes of prolonged blocking events and diagnose model deficiencies. We demonstrate this approach on an idealized quasi-geostrophic model developed by Marshall and Molteni (1993). We train a convolutional neural network (CNN), and subsequently, build a sparse predictive model for the persistence of Atlantic blocking, conditioned on an initial high-pressure anomaly. Shapley Additive ExPlanation (SHAP) analysis reveals that high-pressure anomalies in the American Southeast and North Atlantic, separated by a trough over Atlantic Canada, contribute significantly to prediction of sustained blocking events in the Atlantic region. This agrees with previous work that identified precursors in the same regions via wave train analysis. When we apply the same CNN to blockings in the ERA5 atmospheric reanalysis, there is insufficient data to accurately predict persistent blocks. We partially overcome this limitation by pre-training the CNN on the plentiful data of the Marshall-Molteni model, and then using Transfer Learning to achieve better predictions than direct training. SHAP analysis before and after transfer learning allows a comparison between the predictive features in the reanalysis and the quasigeostrophic model, quantifying dynamical biases in the idealized model. This work demonstrates the potential for machine learning methods to extract meaningful precursors of extreme weather events and achieve better prediction using limited observational data.

Plain Language Summary

Blocking events are an important cause of extreme weather, especially long-lasting blocking events that trap weather systems in place. The duration of blocking events is, however, systematically underestimated in climate models. Using data generated by a simplified atmospheric model we demonstrate that, given sufficient training data, convolutional neural networks can predict the maintenance of Atlantic blocking from an initial blocked state. Next, we show that first training the neural network on data from the simplified model and then fine tuning the training using real world weather data enables prediction even with few examples of long-lasting blocking events in the observational record. Subsequent feature analysis of the resulting neural networks identifies the input variables that most strongly impact their predictions, revealing that areas of high pressure in certain parts of North America and the North Atlantic Ocean are important for predicting long-lasting blocking events and quantifying biases in the idealized model relative to real weather.

1 Introduction

Blocking events are high-amplitude, quasi-stationary anticyclonic high-pressure anomalies that give rise to prolonged abnormal weather conditions in the mid-to-high latitudes (Rex, 1950; Woollings et al., 2018; Lupo, 2021). Blocking events can lead to regional extreme weather by disrupting the usual westerly flow for extended periods (e.g., Kautz et al., 2022), causing extreme heatwaves, floods, and winter storms (e.g., Lupo et al., 2012).

The predictive skill of numerical weather models has improved dramatically, but they still cannot accurately forecast important aspects of blocking events. Blocking frequency and duration are generally simulated poorly by climate models (Davini & D'Andrea, 2020), and even by numerical weather prediction models in medium-range forecasts (Matsueda, 2009; Ferranti et al., 2015; Woollings et al., 2018). Several possible contributing factors have been proposed, including the accuracy of the model's mean flow (Scaife et al., 2010) or synoptic eddies (Berckmans et al., 2013; Zappa et al., 2014a), the model's resolution (Davini & D'Andrea, 2016) and subgrid-scale parameterizations (d'Andrea et al., 1998), and even the choice of blocking index itself (Dole & Gordon, 1983; Tibaldi & Molteni, 1990; Pelly & Hoskins, 2003).

Two commonly used blocking indices (Dole & Gordon, 1983; Tibaldi & Molteni, 1990) highlight two essential features of a blocking *event*: (i) a large positive anomaly of geopotential height that displaces the midlatitude jet, "blocking" the flow, that (ii) persists for longer than typical synoptic variability. Often a five-day (5d) threshold is invoked, but the longer the flow remains in a blocked state, the more severe the implications, either for extended cold/hot conditions or an increased likelihood of compound storm events (i.e., back-to-back storms, which can dramatically increase the potential for damage; Kautz et al., 2022). The persistence of blocking is the focus of our study: given the onset of a blocked state, what is the likelihood that the flow will remain blocked for an extended period, 5 days for a standard event, or up to 9 days for more extreme cases? We take a data-driven approach, training a convolutional neural network to identify persistent blocks at the onset of a blocked state.

To understand blocking, various low-order models have been formulated to identify essential features. In an influential early work, Charney and DeVore (1979) modeled blocking as one of two equilibrium states of a set of dynamical equations for a highly truncated barotropic channel model. Others used low-order models to propose that the positive feedback of synoptic-scale eddies on the blocking structure contributes to the long-time maintenance of blocks (McWilliams, 1980; Hoskins et al., 1983; Shutts, 1983). While these low-order models have provided useful physical insight, their application to the real world is limited by lack of land-sea interactions, topography, and other factors. Comprehensive models, on the other hand, are becoming skillful in simulating realistic blocking [(e.g. Davini et al. (2021))], but their complexity makes it challenging to isolate the essential mechanism(s), and expensive to simulate numerous events.

To strike a balance between complexity, transparency, and statistical robustness from abundant data (model output), we begin with the Marshall-Molteni (MM) model (Marshall & Molteni, 1993), a three-layer quasigeostrophic (QG) approximation of the atmosphere that has previously been used to study blocking events (e.g., Lucarini & Gritsun, 2020). The MM model captures the main features of the northern hemisphere atmosphere reasonably well. For example, Michelangeli and Vautard (1998) found that an enhanced baroclinic wavetrain traveling across the North Atlantic is necessary to trigger the onset of the Euro-Atlantic blocking in both this simple model and reanalysis. They also pointed out that wave-wave interactions and wave-mean interactions dominate local amplification and the propagation of anomalies, respectively.

The MM model allows us the freedom to develop and test methods in a data-rich setting, and precisely quantify the degradation of skill as we pass to a more realistic, data-poor setting. For the particular application of blocking, here we address the question: how well can a data-driven method identify persistent events as a function of the input data you allow it? Furthermore, to gain insight into the physics and predictability of blocking, we turn to Explainable Artificial Intelligence (XAI) techniques, following work by Labe and Barnes (2021) and Rampal et al. (2022). Specifically, we employ Shapley Additive ExPlanation (SHAP) analysis to identify key regions upstream of the blocking center that enable prediction, and use this to construct low-order models the can be interpreted in the context of prior work.

Our ultimate goal, however, is to forecast and understand the maintenance of blocks in our atmosphere, for which we shift the focus to ERA5 reanalysis (Hersbach et al., 2020). For the most extreme case of a 9-day block in the North Atlantic, only 18 have occurred in the historical record (See Tab. 3). What chance does a data-driven approach have? To address the problem of limited data, we apply transfer learning: first we train a convolutional neural network on the MM model to learn the basic features of blocking, and then we re-train it on the limited ERA5 data to calibrate it for the real atmosphere. In this direction our results serve as proof-of-concept. It is likely another choice of physical model could strike a better balance between accuracy and simulation cost for our purpose. Nonetheless, we find that pre-training on the MM model yields a better predictor than when we train the same network on ERA5 alone, proving the efficacy of the transfer learning approach.

The remainder of this paper is organized as follows. Section 2 introduces the Marshall-Molteni (MM) model, training data and blocking index. Section 3 formulates the blocking event criteria and forecasting problem. Section 4 discusses our convolutional neural network structure and training details. We first focus exclusively on the MM model in sections 5 and 6, applying XAI techniques to visualize the important features for prediction and testing the results by building a sparse model with features guided by the XAI. We also suggest physical interpretations for these predictive features. Finally, we turn to the ERA5 data set in Section 7, applying transfer learning to improve the prediction of persistent blocks in ERA5, especially for more extreme events. SHAP analysis shows how transfer learning has modified the CNN to adapt to the new data set, but preserves the use of key upstream regions for prediction.

2 Model and blocking index

Marshall and Molteni (1993) developed a 3-layer quasi-geostropic model of the atmosphere to study atmospheric low-frequency variability. We refer the reader to Appendix Appendix A for a complete description. We use a Northern Hemisphere only version of the model developed by Lucarini and Gritsun (2020) with 6210 degrees of freedom. The model is run with T31 horizontal resolution (corresponding to 90 longitude $\times$ 23 latitude gridpoints across the northern hemisphere). All model output fields, as well as the reanalysis used later, are averaged daily.

We use an index developed by Dole and Gordon (1983) to define blocking events, hearafter referred to as the DG index. This is an anomaly-based blocking index, but has been shown to capture the same essential features of blocking as other measures, e.g., that of Tibaldi and Molteni (1990). We compute this index by transforming the spherical harmonic representation of the streamfunction ψ at 500 hPa into approximate geopotential height, Z, on a Gaussian grid for latitude and a uniform grid for longitude. The approximation is the choice of a fixed Coriolis parameter f_0 to convert from ψ to Z, which causes minimal distortion over our midlatitude area of focus. Blocks are based on deviations of the geopotential height from climatology, denoted Z'.

A *blocking event* is said to occur at a specific location when Z' stays above a tunable geopotential height anomaly threshold, M, for at least five consecutive days. In their paper, Dole and Gordon (1983) tested statistics for varying M values, ranging from 50 m to 250 m, with subsequent studies adopting different thresholds (Chan et al., 2019, Tab. 2). For our investigation, we calibrated $M = 100$ m for our MM model simulation to roughly match the blocking fraction computed from ERA5 reanalysis data, where we used the threshold $M = 150$ m as in Mullen (1987).

Fig. 1 shows the blocking event statistics during the simulation. For comparison, blocking event statistics computed from ERA5 reanalysis data from 1959-2021 are also shown. In this study, we focus on North Atlantic blockings indicated by the white rectangle in Fig. 1. We pick this region because it has a relatively high blocking frequency, and for its important influence on western Europe. We use Z_B, the average 500 hPa geopotential height anomaly over this target region over the North Atlantic, to define blocked states and blocking events.

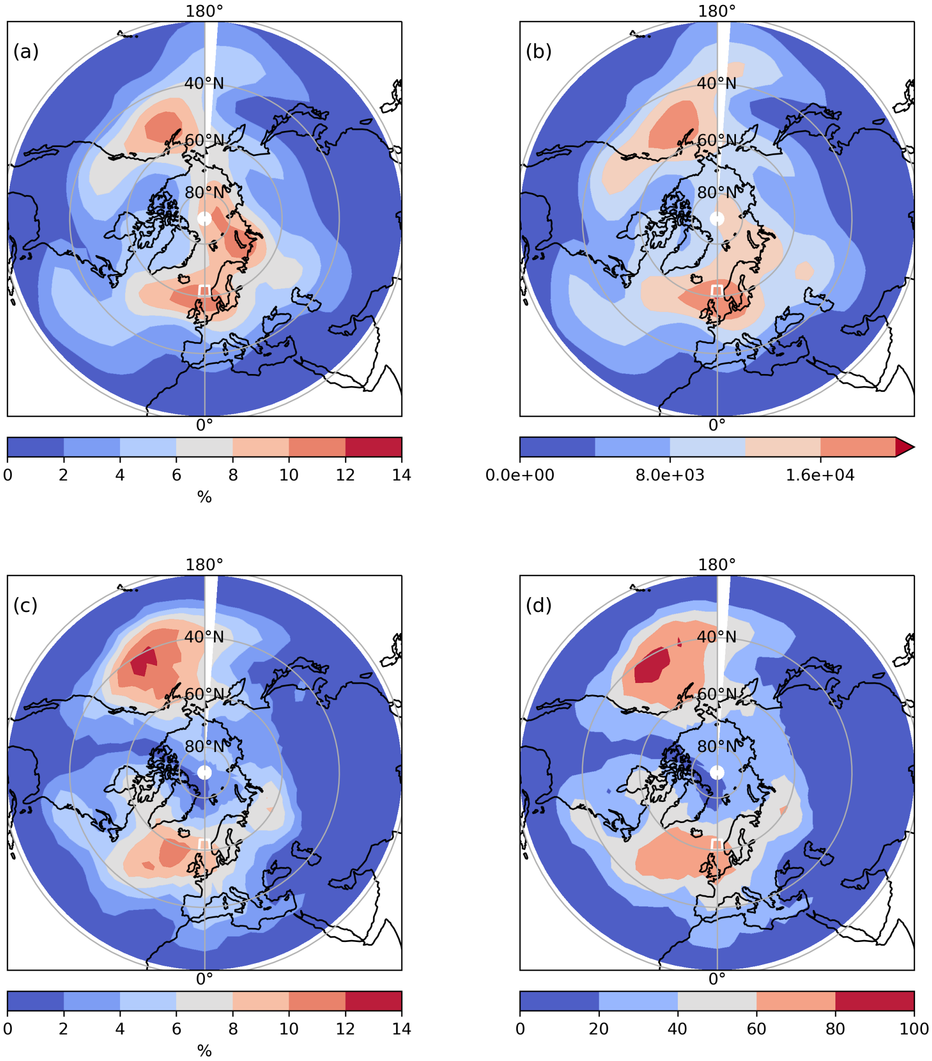

Figure 1. (a) Blocking fraction (the percent of days with $T \geq 5$ days) for MM model data with $M = 100$ m. (b) total blocking event counts for MM model data during the simulation. (c) blocking fraction for ERA5 reanalysis data with $M = 150$. (d) total blocking event for ERA5 reanalysis data with $M = 150$m. In all subfigures, the region we focus on is indicated by the white rectangle centered at 0°E and 62°N (approximately spanned by 3 longitude points covering 4°W-4°E, and 2 latitude points covering 60°N-64°N)

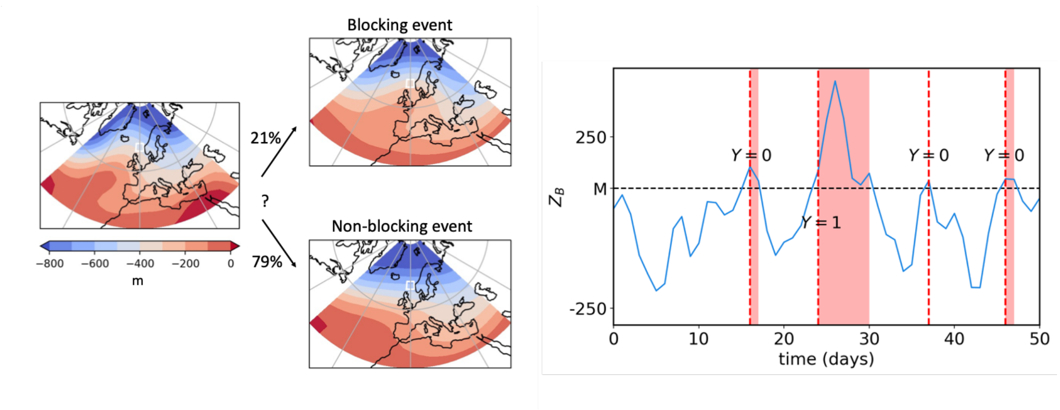

Figure 2. Left: The blocking persistence problem: given a nascent blocked state, the goal is to forecast whether it will persist into a long-lasting blocking event, or quickly return to climatology. The percentile represents the climatological probability. Right: A sample trajectory of $Z_B(t)$, the anomaly of geopotential height defined in Sec. 2. The vertical dashed lines indicate new blocked states ($T = 1$). The red shading indicates the duration of the block. The label $Y = 1$ indicates that the blocked state persisted 5 days to constitute a blocking event, while $Y = 0$ indicates that it did not.

3 Probabilistic forecasting and event definition

We aim to study the *maintenance* of blocks rather than their *onset*. Precisely, we formulate the question as the classification problem posed in Fig. 2: given a nascent blocked state, i.e., the state on a day that geopotential height anomalies over the North Atlantic first exceed the threshold M, can we immediately predict whether the flow will remain blocked for 5 or more days—evolving into a *blocking event*—or will the flow return back towards the climatological state before 5 days have passed? In the MM model, nascent blocked states evolve into 5-day persistent blocking events approximately 21% of the time.

We pose the classification problem: given only the state at the time of blocking onset, can a data-driven method accurately identify the rarer cases that will persist for more than 5 consecutive days? Mathematically, we denote the full model state by $\boldsymbol{X}$ and further introduce a variable T for the running duration of a blocked state:

$$T = (\text{days since } Z_B < M). \tag{1}$$

Note that $Z_B(t)$ is determined by the state vector $\boldsymbol{X}(t)$ at any time t, but $T(t)$ retains some memory of previous states and thus is not fully determined by $\boldsymbol{X}(t)$. For example, as shown in Fig. 2, suppose $Z_B(t)$ first rises above M on day $t = 16$ and dips back below M on day $t = 18$. Then, $T(t) = 0$ for all days through $t = 15$, $T(16) = 1$, $T(17) = 2$, and $T(18) = 0$. With this notation, we can say that "$\boldsymbol{X}(t)$ is the beginning of a blocking event" if

$$T(t) = 1 \quad \text{and} \quad T(t + D - 1) = D. \tag{2}$$

The condition $T(t+D-1) = D$ only holds when there are at least D consecutive days with $Z_B(t) \geq M$ starting from t. We can see an example of this in Fig. 2 at day 24, for both a block of duration 5 and 7 d. Here, $T(24) = 1$, and $T(28) = 5$, triggering the condition for $D = 5$. The flow remains blocked through $T(30) = 7$, such that day 24 would also count as the onset of a $D = 7$ day blocking event.

With this formulation, our central question becomes: given a $T(t) = 1$ state at time t (the flow has just become blocked), will it stay blocked for D days, $T(t + D -$

$1) = D$, or not? We address this question by estimating the conditional probability:

$$q(\boldsymbol{x}(t)) = \mathbb{P}[T(t+D-1) = D \mid \boldsymbol{X}(t) = \boldsymbol{x}(t), T(t) = 1]. \tag{3}$$

Many recent studies have summarized extreme climate and weather events via similar functions of state (e.g., Tantet et al., 2015; Finkel et al., 2021; Lucente et al., 2022; Jacques-Dumas et al., 2023; Miloshevich et al., 2023). Unless otherwise specified, we adopt $D = 5$ to maintain consistency with the common blocking indices (Dole & Gordon, 1983; Tibaldi & Molteni, 1990; Pelly & Hoskins, 2003). We also consider more extreme events with $D = 7$ and $D = 9$.

4 Convolutional Neural Network Training and Performance

Convolutional Neural Networks (CNN) have gained widespread application in probabilistic forecasting problems (Liu et al., 2016; Ham et al., 2019; Miloshevich et al., 2023) for their outstanding performance on multidimensional data sets with spatial structure. A CNN differs from a dense neural network in the use of convolutional layers with shared weights and biases across layers within the network, designed to extract features that exhibit translation invariance across the input space (Goodfellow et al., 2016). Originally developed in the context of image processing, CNN excels in scenarios where target objects, such as the face of a cat, may appear at different places within the training image. Convolutional layers allow the network to efficiently learn predictive features, combining information across multiple images. In our context, we expect predictive contributions from atmospheric eddies and Rossby waves, which share similar dynamics across all longitudes. A CNN can potentially extract these features more effectively than a fully connected architecture could, while still learning how they vary with longitude due to topography and other zonal asymmetries.

The structure of the CNN in this investigation follows Miloshevich et al. (2023) and is shown in Fig. 3. It consists of a three-layer architecture, combining convolutional filters followed by ReLu activations. Specifically, we use 32 and 64 filters (3×3) for the first and last two convolutional layers. Between each pair of convolutional layers is a max-pooling layer. The output is then flattened and passed to a dense layer with 64 neurons that produces 2 outputs. Finally, a softmax function converts these two outputs to complementary probabilities.

We performed experiments with alternative CNN structures and found that reducing the widths of layers mitigates overfitting, but also reduces the performance at the best epoch (not shown). Therefore we adopt the architecture in Fig. 3 and use early-stopping to avoid overfitting, as detailed below.

4.1 Training and Test Datasets

We create a training and test set of all states where the flow has just become blocked: $\{(\boldsymbol{X}, T) | T = 1\}$, where $\boldsymbol{X}$ are $18 \times 90 \times 3$ (latitudes $\times$ longitudes $\times$ pressure at levels of 200 hPa, 500 hPa, 800 hPa) grid maps of geopotential height from 20°N to 87°N. Our goal is to classify which of these cases persist into blocking events ($Y = 1$) versus states that do not ($Y = 0$). Fig. 2 shows a sample time series with 4 instances of a nascent blocked state, $t = 16, 24, 38$ and 47, only the second of which evolves into a persistent blocking event: $Y = 0$, 1, 0, and 0, respectively. For each case, the model must classify $Y = 0$ or $Y = 1$ given only $\boldsymbol{X}$ at the onset time.

We examined the sensitivity of CNN model performance with respect to different amounts of training data. To prepare the dataset, we integrate the MM model for 1250k days in total. The computational cost is low, requiring 1 CPU core and approximately 11 hours. We select the first n days (with n ranging from 1k to 1000k) to create the training data set, and always take the last 250k days for the test dataset. Thus all models

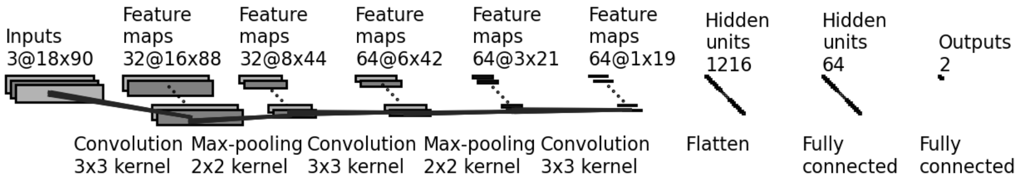

Figure 3. Convolutional Neural Network structure. The three convolutional layers respectively use 32, 64 and 64 filters (3×3), followed by ReLu activations. Between each pair of convolutional layers is a max-pooling layer with window size 2×2. Then the output is flattened and passed to a dense layer with 64 neurons that produces 2 outputs. A softmax function maps these outputs to two positive numbers between zero and one, representing the estimated probabilities of the the nascent blocked state to persist or decay.

Training data		Test data	
Days	Nascent blocked states	Days	Nascent blocked states
1k	63		
10k	699		
100k	7024	250k	17755
500k	35078		
1000k	70635		

Table 1. Length of trajectory (in thousands of days) vs. number of nascent blocking states ($T = 1$) in training set and test sets of varying size.

can be fairly compared. The trajectory length and the corresponding number of nascent blocked state states are shown in Tab. 1. The likelihood q of forming a blocking event varies depending on different persistence thresholds D. This dependence relationship is illustrated in Tab. 2.

4.2 Learning procedure

For simplicity, we use binary cross entropy as a loss function, a common choice for classification (Miloshevich et al., 2023). Alternative loss functions have been studied by Rudy and Sapsis (2023). The loss function $L(q)$ is defined as as follows:

$$L(q) = -\frac{1}{N}\sum_{i=1}^{N}\Big[Y_i \log q(Y_i = 1) + (1 - Y_i)\log(1 - q(Y_i = 1))\Big]$$

where $q(Y_i = 1) \in (0, 1)$ is the probability of the event $Y_i = 1$ as predicted by the CNN. $L(q)$ is small when the CNN assigns high probability to positive events and low probability to negative events.

Given the rarity of blocking events, the data exhibit a pronounced class-imbalance, which becomes increasingly severe for longer block durations. As shown in Tab. 2, for $D = 5$, only about 1 in 5 nascent blocked states persist into an event, but $D = 9$, less than 1 in 20 evolve into persistent events. With this extreme imbalance, a model that

Threshold	$Y = 1$	$Y = 0$	Positive rate
≥ 5 d	18748	69642	0.212
≥ 7 d	8522	79868	0.096
≥ 9 d	3891	84499	0.044

Table 2. The statistics of blocking events in our MM 1250k day simulation. The full dataset exhibits 88390 nascent blocking states ($T = 1$ states). $Y = 1$ marks the number of these nascent blocks that persist for 5, 7, or 9 d, thus evolving into a blocking event under these respective thresholds, while $Y = 0$ denotes the number that don't make it to the threshold.

never predicts an event will be correct over 80% or 95% of the time, respectively. However, such a model would clearly underperform in terms of precision and recall (defined in the next subsection), which would both be zero.

To address the class imbalance, for our results in this section we employ over-sampling (Johnson & Khoshgoftaar, 2019) techniques during training. In each epoch, we sample an equal number of nascent blocks from both classes until we complete an iteration over all the nascent blocks in the overrepresented class. As a result, the nascent blocks that persist have been sampled multiple times during each epoch.

4.3 Performance metrics

Throughout this study, we evaluate model performance using two key metrics: *precision* and *recall*. We monitor the values of these metrics on the test dataset throughout the training process to determine the stopping point in order to avoid overfitting. The precision and recall are respectively defined as

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}, \tag{4}$$

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}, \tag{5}$$

where "True positives" is the number of data points with $Y = 1$ for which our CNN correctly predicts a persistent blocking event; "False positives" is the number of data points with $Y = 0$ for which our CNN incorrectly predicts a persistent blocking event; and "False negatives" is the number of data points with $Y = 1$ for which our CNN incorrectly predicts that the blocked state does not persist.

More informally, the precision measures the fraction of *forecasted* persistent blocks that *actually* persist. The recall, on the other hand, is the fraction of *actually* persistent blocks that are successfully *forecasted*. If one randomly predicts events with the climatological mean rate p, regardless of the system state, then the precision and recall are both given by $\frac{p^2 N}{p^2 N+(1-p)pN} = p$. This sets the floor for a useful predictor: both the precision and recall must be higher than the climatological rate.

There can be tradeoffs between improving the precision and recall. Predicting the event all the time will give you a perfect recall, but climatological precision p. A low recall implies missing a substantial number of positive events, leading to inadequate preparation and increased risk of damage. Conversely, a low precision suggests over-predicting events, "crying wolf" too often. In the context of extreme weather forecasting, this can lead to over-preparation, consequently reducing the efficiency of regular societal operations, as well as trust.

A reasonably high value of both recall and precision is crucial for an effective and resource-efficient forecasting model. We use a simplistic definition of 'best' performance, expressed as

$$\text{Overall performance} = \text{Precision} + \text{Recall}\,. \tag{6}$$

However, it is crucial to note that in practical scenarios, designing overall performance metrics requires careful consideration of the cost of preparing vs. risk of damage associated without preparation.

This naïve criterion only works when the precision and recall are both reasonably high, since forecasting the event all the time will yield a performance score of $1+p$ (recall of 1 and precision of p). We used caution in ERA5 based forecasts, requiring our trained models exhibit nontrivial precision above the climatological rate. We found that the F1-score (Sasaki et al., 2007), another common performance metric, selects the same epoch as the metric in (6).

4.4 Performance and early stopping technique

The top row of Fig. 4 shows the precision and recall evaluated on the test data for varying training data sets for $D = 5$. Both the precision and recall metrics are plotted starting from the end of Epoch 1 (the leftmost point on the horizontal axis of Fig. 4); From Epoch 2 to Epoch 10, the precision increases, chiefly reflecting a decrease in the false positive rate, as the CNN becomes better at discriminating between persistent and non-persistent flow configurations. At the same time, the recall slowly decays: the false negative rate rises slightly as the network becomes more conservative and less likely to over-predicting persistent cases. Except for the low data regime (1k days), the performance of the CNN asymptotes after approximately 10 epochs where the precision and recall are approximately equal, but this is not necessarily the ideal stopping time (Miloshevich et al., 2023).

To select the CNN parameters with the best performance, we assessed the overall performance defined in Eq. (6) at the end of each epoch. We then use the parameters from the epoch with the largest value. The "best" CNN is obtained by training on the full data set of 1000k days for 4 epochs, indicated by the star in Fig. 4. It achieves precision of 0.70 and recall of 0.87, exhibiting significant predictive power over the climatological mean prediction (the black dashed line with value 0.21). Therefore, we use it for further analysis in Sec. 5.

All of our CNNs significantly outperformed the climatological mean prediction for any amount of data or training length. Interestingly, although the best performance is always realized with the longest trajectory of 1000k days, precision and recall have different sensitivities to the training data size. For $D = 5$ events, the precision improves with more data up to 100k days (equivalent to approximately 1000 winters), after which additional data does not lead to much improvement. The recall, however, is more data-hungry; its performance continues to improve with more data up to 500k days, equivalent to 5 millennia of winter data. This reflects the fact that more data continues to help the CNN avoid missing events after its ability to limit false positive forecasts has saturated.

Fig. 4 also shows the results for higher persistence thresholds, $D = 7$ and 9. These thresholds correspond to rarer events, and even with the longest trajectory of 1000k days, the precision and recall curves suffer for two reasons. First, as seen from Tab. 2, the number of positive events drops, effectively limiting the data set almost by a factor of 5 for the most extreme $D = 9$ cases. More importantly, however, it simply becomes harder to discriminate rare events as the data set becomes more imbalanced: less than 1 in 10 nascent blocking states will evolve into a 7 d blocking event, and less than 1 in 20 into

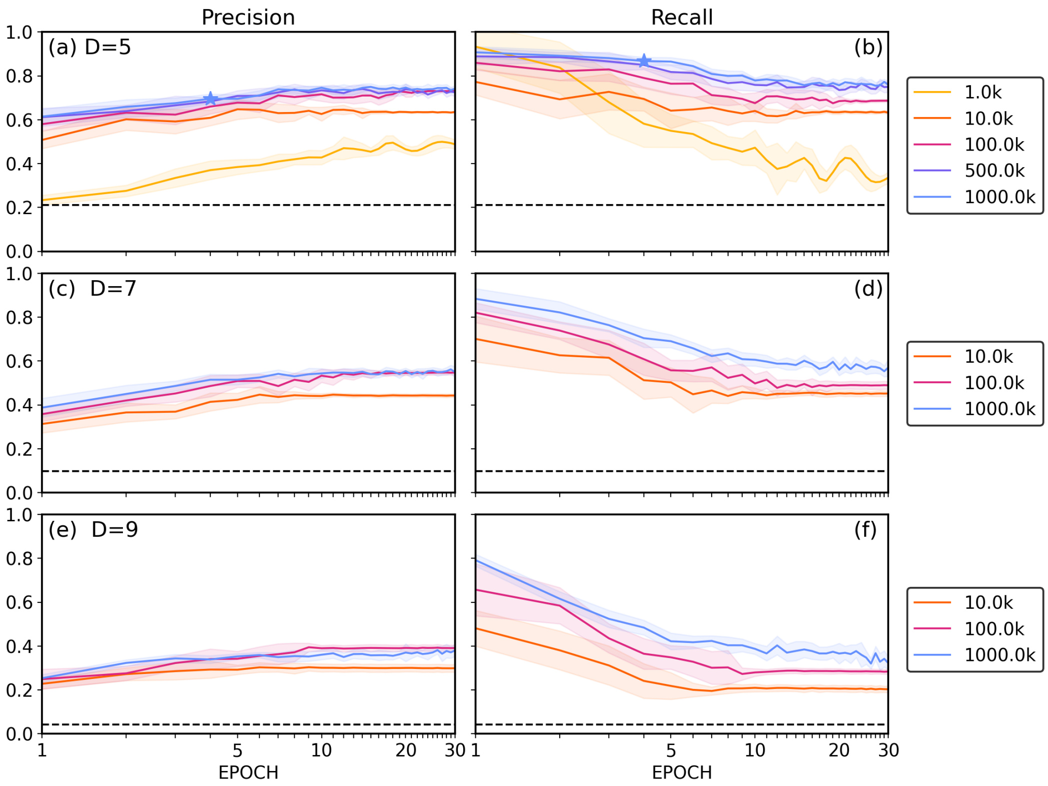

Figure 4. Precision (a,c,e) and recall (b,d,f) as a function of training epoch, for CNNs trained on datasets of varying sizes (curve color) and thresholds of blocking persistence (rows). As detailed in the text, all the models are tested on events from the same 250K dataset not seen in training. Panels (a,b) show results for 5 day blocks (D=5); for example, the light blue curves are trained on all events in the full 1000k-day simulation, while the other curves show results based on smaller training sets as indicated by the legend. The blue stars indicate the "best" CNN (see text), with a precision=0.70 and recall=0.87. Panels (c,d) show results for D=7 blocks and (e,f) for D=9 blocks. Fewer curves are displayed for $D = 7$ and $D = 9$ for the sake of clarity. Shading indicates uncertainty, assessed by taking one standard-deviation of results of ten neural network training with i.i.d random parameter initialization.

a 9 d blocking event. Without our efforts to overcome this imbalance, a network can classify almost all events correctly by never predicting a persistent case.

Despite the difficulties, the CNNs still show some skill in rare event forecasting. Given the full 1000k dataset, for $D = 9$ the precision and recall converge to about 0.35, a factor of two worse than the CNN in the $D = 5$ case but a factor of 10 better than climatology. As with the D=5 cases, we found that the recall for $D = 7$ and 9 suffers more than the precision when the data set shrinks: with less events to learn from, the CNNs become more conservative and less likely to call an event. The recall depends on the false negative rate, and thus appears more sensitive to class imbalance. More data gives the network more true positive cases to learn from, apparently helping to overcome this challenge.

The low precision and recall values for smaller data sets (1k and 10k) do not bode well for training our CNN on ERA5 data, which will be discussed in detail in Section 7. For $D = 5$, there are 273 nascent blocked states in the ERA5 record, 84 of which persist into blocking events (see Table 3). This data amount falls between our 1k and 10k cases where data clearly limit performance. Consistent with our experience with the MM model, recall is the metric that suffers most from limited data, and stands to benefit the most from transfer learning.

5 Feature analysis: What is our CNN using to predict blocking events?

Before turning to forecasting in the realistic data regime, we ask what our best CNNs have learned to make these forecasts. Explainable Artificial Intelligence (XAI) is an array of techniques used to try to gain some understanding of the basis on which neural networks make predictions (Linardatos et al., 2020). In this section, we use SHapley Additive exPlanation (SHAP) value analysis to dissect the contributions of different atmospheric pressure levels and geographic areas that our CNN is using to make its predictions. We further construct a sparse model using the identified important features as inputs to quantitatively justify their relative importance in the prediction process.

5.1 XAI Method

SHapley Additive exPlanation (SHAP) values, introduced by Lundberg and Lee (2017) and Shrikumar et al. (2017), draw inspiration from Shapley values in game theory (Lipovetsky & Conklin, 2001). In the domain of weather and climate science, SHAP values have found broad use, with applications ranging from Earth System model error characterization (Silva et al., 2022) to drought forecasting (Dikshit & Pradhan, 2021).

Intuitively, given a function $f : \mathbb{R}^d \to \mathbb{R}$ (such as the conditional probability q in Eq. 3), SHAP assigns an importance value ϕ_i to each feature x_i of the argument $\boldsymbol{x} \in \mathbb{R}^d$, which combine additively:

$$f(\boldsymbol{x}) = \mathbb{E}[f(\boldsymbol{x})] + \sum_{i=1}^{d} \phi_i(f, \boldsymbol{x}). \tag{7}$$

With no knowledge of $\boldsymbol{x}$, the optimal prediction of f (in a mean-square sense) is the climatological average over the distribution of $\boldsymbol{x}$: $\mathbb{E}[f(\boldsymbol{x})]$. SHAP values quantify how much is gained beyond this baseline by incorporating information from each component i of $\boldsymbol{x}$. The SHAP values $\phi_i(f, \boldsymbol{x})$ are unique for each sample of $\boldsymbol{x}$, but features i for which $|\phi_i(f, \boldsymbol{x})|$ are large for most $\boldsymbol{x}$ (that is, a large average SHAP value) can be singled out as important, or useful, for the prediction of $f(\boldsymbol{x})$. SHAP values possess advantageous theoretical properties as well, and we refer the reader to Lundberg and Lee (2017) for a detailed theoretical analysis. In this study, SHAP values are computed using the Python package `DeepSHAP` (Chen, 2022). The function $f(\boldsymbol{x})$ is taken as the estimated conditional

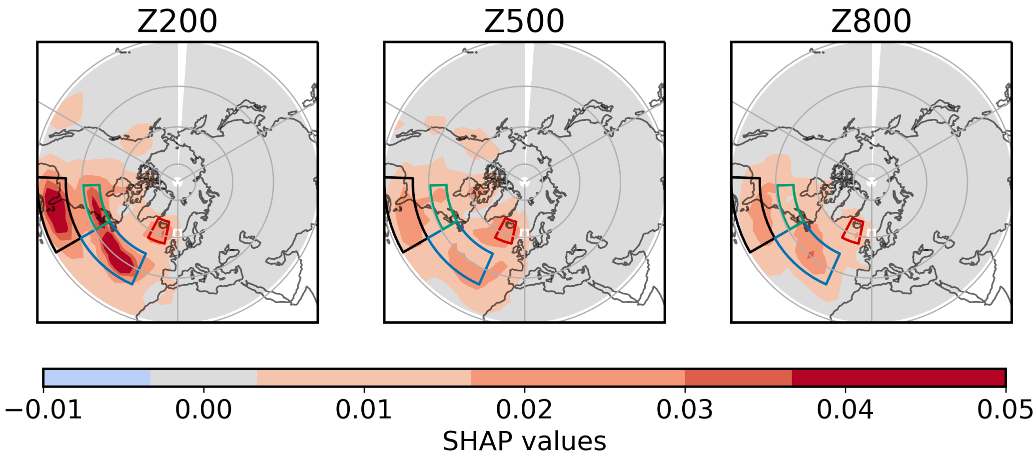

Figure 5. Composite maps of SHAP values, $\overline{\phi}$, of geopotential height at 200, 500, and 800 hPa, for true positive cases, i.e., when the CNN accurately forecasts a persistent blocking event. The unit is the probability per feature (Z at a given location and pressure level) of a positive forecast (see equation 7), indicating the feature's average incremental contribution to the CNN's confidence that the nascent blocked state will evolve into a persistent blocking. The boundaries of the most important regions learned by the CNN are marked by solid lines and denoted region 1 (Florida, black), region 2 (north Atlantic, blue), region 3 (northeastern North America, green) and region 4 (Iceland, red) .

probability $\hat{q}(\boldsymbol{x})$ computed by the CNN, i.e., the probability, according to the CNN, that the blocked state will extend $\geq D$ days, leading to a blocking event.

5.2 Results

Fig. 5 shows the composite of SHAP values for true positive data. Because few nascent blocks persist for $D = 5$, 7, or 9, the climalogical probability of a persistent event $\mathbb{E}[\hat{q}(\boldsymbol{x})] =$ 0.21, 0.096, and 0.044, respectively. For our CNN to call a positive event, we require the conditional forecast probability $\hat{q}(\boldsymbol{x})$ to be larger than 0.5. Hence a positive (negative) value of $\phi_i(\hat{q}, \boldsymbol{x})$ indicates that knowing the geopotential height anomaly at this level and location increases (decreases) the likelihood of a positive event. Therefore, the shading in Fig. 5 can be interpreted as the average influence of each grid point for the CNN to successfully predict a long-lasting blocking event. For the averages over each region, the standard deviations for Z200, Z500, and Z800 are 0.039, 0.026, and 0.028, respectively, with a roughly symmetric distribution, indicating that the SHAP value analysis in Fig. 5 represents the overall sample behavior, rather than being skewed by outliers.

The SHAP composite is approximately uniformly non-negative because it is based only on true positive events: additional information should always increase the forecast probability. This indicates that the CNN has been well-trained to only use geopotential height information that improves the blocking event probability, and suggests it has identified robust features that herald a persistent block. A composite based on true negative cases (not shown) reveals similar patterns, but of the opposite sign.

The first thing to notice is that anomalies upstream from the blocking region (to the west) are more valuable than other regions for predicting the persistence of the blocked state. Moreover, the commonality among different pressure levels reflects the relatively barotropic nature of the MM model. In general, however, the CNN prediction relies most on the upper level flow (200 hPa).

The SHAP values emphasize four distinct regions in a quadrupole arrangement to the west of the Atlantic blocking region, as marked in Fig. 5. We chose these regions to encapsulate high SHAP values using the following algorithm: after objectively identifying regions where SHAP values exceeded a set threshold, we defined boundaries by hand with the goal of enclosing these regions across all three levels within the smallest encompassing rectangle. While part of the goal of choosing these regions was to build a sparse predictor in the next section, they give us physical insight on their own.

The meaning of the SHAP values can be more easily interpreted with the aid of composites of the 3341 true positive events (Fig. 6), which show us the sign of anomalies that favor persistence. Positive geopotential anomalies in region 1 (black, centered over Florida) and 4 (red, over Iceland, just east of the blocking region itself) at the onset of blocking indicate to the CNN that a block will persist, while negative anomalies over Regions 2 (blue, North Atlantic Ocean) and 3 (green, northeast US) also favor persistence.

Regions 2 and 4 project onto opposing centers of action of the North Atlantic Oscillation (NAO). They indicate that a more negative NAO state at the onset of blocking increases the likelihood of a persistent block. Previous studies have also found that blocks tend to be more persistent when the NAO is negative (Barnes & Hartmann, 2010). While a blocking pattern off Europe projects weakly onto the NAO itself, SHAP analysis indicates that the wider structure of the pattern is important. Regions 1, 3, and 4, on the other hand, appear to be part of a wave train arching southwest from the blocking region. Their importance suggests that downstream development of a wave packet propagating along the jet stream helps drive persistent blocking events in the North Atlantic.

6 Building a sparse model with logistic regression

In quantifying the relative importance of the geopotential height as a function of location, the SHAP values suggest there is potential for dimension reduction. The CNN did not rely significantly on the information about Z in the large grey regions downstream of our target blocking area in Fig. 5 to predict the potential persistence of a nascent blocking anomaly. Intuitively, conditions over central Asia will take some time to affect the flow over the North Atlantic, and are mostly irrelevant for a forecast at 5 day range.

To gain more physical insight into the utility of the SHAP values, and so gain confidence in our CNNs, we explored a simplistic dimension reduction approach focused on the regions highlighted in Fig. 5. Our aim was not to achieve the ideal dimension reduction, but to provide physical insight. Thus we ask: how well can one predict the persistence of a nascent block given only the very coarse information about the flow provided by the average geopotential height within these regions at the three levels?

For these simple models, we computed the local mean of $Z200$, $Z500$, and $Z800$ for each of the four rectangles shown in Fig. 5, resulting in 12 time series. We then applied logistic regression with all possible combinations of these 12 features for subsets of dimension up to 5, i.e., for dimension 1, fitting a logistic function with each time series alone, for dimension 2, all possible combinations of two time series, and so forth. The results for the sparse models with the best predictive skill on the test set are illustrated in Fig. 7(a). The horizontal axis denotes the combinations of variables that achieve the predictive scores shown in the figure.

We draw three key conclusions from Fig. 7(a). First, to predict the persistence of a blocked state, the best one-dimensional feature is Z200 in region 1, over Florida and

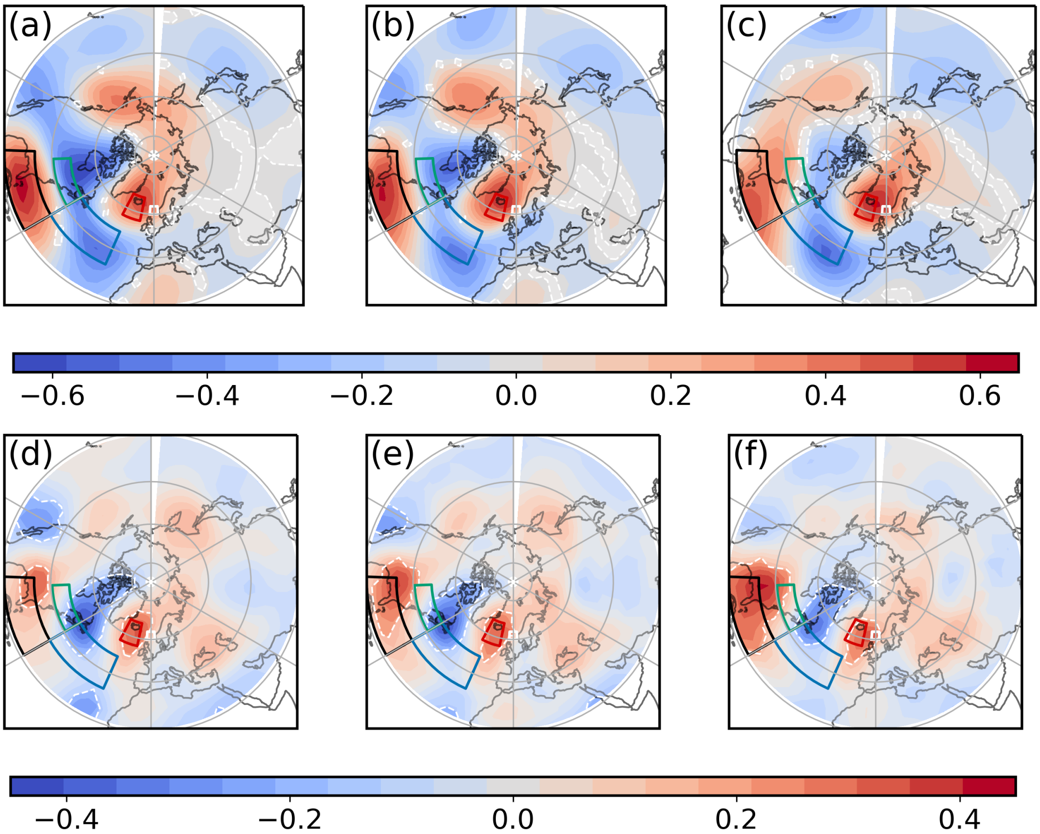

Figure 6. Averages of nascent blocking states that evolve into persistent blocking events ($T = 1, y = 1$) of (top row, a-c) MM dataset and (bottom row, d-f) ERA5. The colorbar represents values of geopotential height anomalies normalized by the standard deviation at each location and height. The white dashed line indicates the 0.05 significance level for a one sample t-test of the null hypothesis that the expected value is zero. The box areas identified by SHAP analysis lie in statistically significant regions. Regions that are not significant are shaded by white. For MM model dataset (the top row), most of the regions are statistically significant, while for ERA5 dataset (the bottom row) most of the regions are not statistically significant.

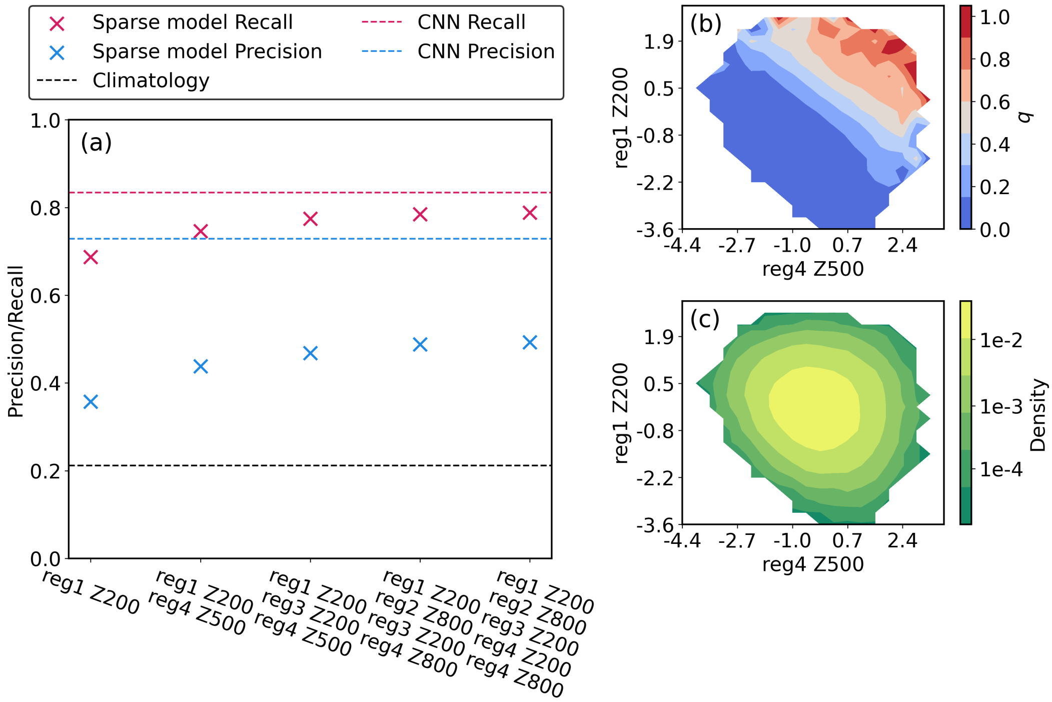

Figure 7. (a): Sparse model predictive skill on the test data set. The horizontal axis represents the dimension d of the sparse model from 1 to 5, with labels showing the combination of variables ("R1" = "region 1") that achieves the best predictive skill among all combinations of d variables. (b) Conditional probability of a persistent block, q, as a function of mean normalized geopotential height anomaly at 200 mb over region 1 and at 500 mb over region 4 (the second column of (a)). (c) The marginal density (likelihood of observing these anomalies) as a function of the same variables. Densities below 10^{-5} are cut off.

the Gulf and upstream of the block, not Z500 in region 4, the $Z-$field nearest to the block. Second, the combination of Z200 in region 1 with Z500 in region 4 forms a two-dimension model (shown in Fig. 7(b)) that already recovers a recall value of 0.75—it captures three quarters of all blocking events—with a precision of 0.44, twice the climatological rate. The precision and recall of the full CNN, however, are 0.87 and 0.70. This leads us to the third key message: there is a large discrepancy in precision between CNN and logistic regression. Even with 5 predictors, the precision of our sparse model is only 0.5.

The poor precision indicates that the sparse model makes too many false positive predictions. This could suggest that the decay of the Atlantic blocked state is a more nonlinear dynamical phenomenon, which cannot be modeled as a simple linear statistical model. A CNN can capture these nonlinearities more effectively than sparse regression, which is consistent with previous research which found North Atlantic blocks are associated with nonlinear processes (Evans & Black, 2003). It could also indicate that more subtle features outside these 4 centers (and variation within these regions) are important. Fig. 5 indicates that the CNN uses information across all of the North Atlantic, eastern North America, and even off the west coast of the US, to make skillful predictions.

To explore the effectiveness of the two-dimensional sparse model, we visualized the conditional probability of a block persisting, q, projected onto this simple subspace (shown in Fig. 7(b)). For example, the lightest pink region, corresponding to $q \approx 0.5$ indicates that if, at the onset of blocking, Z at 200 hPa over region 1 (Florida) is particularly high or Z at 500 hPa in region 4 (Iceland) is abnormally high, the system has a roughly 50% chance of evolving into a persistent block, more than double the climatological rate of 21%. In the red region at the top right, where both of these regions exhibit abnormally high pressure, the chance of a persistent block increases to near 100%.

Fig. 7(c) shows the likelihood of observing these Z200 and 500 anomalies. Most often, the system exists in the middle of the diagram, where the probability of a blocking event hovers near or below the climatological value. The most likely state that exhibits a high chance of a block lies along the diagonal from the upper left to the lower right, with moderately high Z200 and 500 anomalies. The states in the top right corner, for which a persistent block is nearly certain, are very rare.

The sparse models suggest physical links between blocking events and the upstream flow. The Atlantic blocking region lies at the end of the Atlantic storm track (Michelangeli & Vautard, 1998). Persistent blocks, at least in the MM model, are favored when there is enhanced wind off the east coast of the US (high pressure over Florida, region 1) and low pressure over regions 2 and 3 (which are highlighted in the higher dimensional sparse models). This displaces the climatological winds upstream of the blocking region equatorward. This will modify the input of storm activity into the blocking region, consistent with prior studies that have highlighted the relation between the storm track and blocking events (Zappa et al., 2014b; Yang et al., 2021).

7 Extending to ERA5 using Transfer Learning

Given sufficient data, it was possible to construct a CNN that skillfully forecasts the maintenance of blocking events in the MM model. However, the ERA5 data from December, January and February (DJF) between 1940-2022 exhibit only 273 nascent blocked states in our Atlantic region of focus. Unfortunately, this low-data regime is where we see a significant degradation in performance in Fig. 4. The curve associated with the trajectory of 10k days (699 nascent blocked states) plateaued at lower values for both the precision and recall. With only 1k days (63 nascent blocked states) performance was poor, and the learning unstable, oscillating significantly across epochs.

The class imbalance between $Y = 0$ and $Y = 1$ adds to the difficulty (see Tab. 3), particularly when longer blocks are considered. An extreme example is the set of blocking events that last $\geq$9 d: there are only 18 such events in the reanalysis record out of 273 data points. Such a small sample of positive data can hardly support any meaningful training, and makes it impractical to get meaningful uncertainty bounds on performance. In a standard training-test data split with a ratio of 90:10, only around 2 positive events typically fall in the test set, making it challenging to robustly assess the skill.

When training on the limited number of events in the reanalysis, a CNN can more easily suffer from overfitting, where the network uses 'noise' (unrelated features) to classify blocking events. Overfitting can be diagnosed when the performance on the test set diverges from the training set. Yang and Gerber (submitted) found that the oversampling strategy used so far in this study was more prone to overfitting than a weighted loss function strategy (Johnson & Khoshgoftaar, 2019). With this latter strategy, one emphasizes the rare class (in our case, positive events) by increasing its weight in the loss function. In our remaining experiments, we weighted positive and negative events inversely to their occurrence rate.

7.1 Direct training

The scarcity of events makes direct training (DT) on ERA5 blocks challenging. In our study of the MM model data, we had the luxury of a large test data set (which we

Threshold	$Y = 1$	$Y = 0$
≥ 5 d	84	189
≥ 7 d	36	237
≥ 9 d	18	255

Table 3. The statistics of ERA5 dataset in 1940-2022 DJF with $T = 1$.

intentionally kept the same for fair comparison of the different CNNs), even for the case with only 1k training days. For ERA5 data, we use cross validation (Goodfellow et al., 2016) to make the best use of the smaller dataset. The limited number of states were partitioned into training and test sets in ratios of 90:10; we also tried 80:20, and the results were similar (not shown). These splits were chosen to balance two difficulties: a small training set can prevent robust learning, while a small test data set limits accurate evaluation, even for a well-trained model.

To proceed, we first reduced the resolution of the ERA5 data to a comparable size of the MM output, considering geopotential height on the same three levels at the same coarse resolution. Reducing the resolution allowed us to use the same CNN architecture, and made transfer learning possible (as discussed below). It also helped avoid overfitting, reducing the number of input variables relative to the number of events. Then we created the test-train splits, yielding 10 cross validation sets with distinct test events. Finally, for each test-train split, we trained and evaluated 10 CNNs, where variations were confined to random weight initialization and shuffling of training data.

Providing meaningful uncertainty on the precision and recall statistics from direct training, shown in the left column of Fig. 8, is challenging. As the 10 CNNs trained on each train-test split are not independent and identically distributed (IID), we first average the skill scores within each split. The 10 test sets, however, can be viewed as IID samples. The solid lines and shades respectively represent the mean and two-standard deviation bounds of the precision and recall, as a function of epoch, across the 10 splits.

For 5 d blocks, a CNN trained by DT can beat the climatological forecast, albeit only modestly. Given the small testing data set (27 nascent blocks, of which roughly 8 persist into events), it is important not to put too much stock in the best possible performing network, for CNN can get lucky on a small sample size. The average performance quantifies the potential skill more reliably. On average, a CNN can achieve a precision of approximately 0.45: when it calls a persistent blocking event, 4-5 out of 10 times it is correct, as compared to about 3 of 10 in the climatology. The recall was modestly better, the network only missing 4 of 10 actual events, while a climatological forecast would miss 7 of 10.

We also explore 7 d events, where only 13% of nascent blocks evolve into 7+ d events. Again, the average CNN modestly beats the climatological forecast in terms of precision: 1/5 of the cases it calls evolve into persistent events, roughly double the success rate by a guess with a Bernoulli random variable. The recall was initially deceptively high (the network captured 5 of 10 blocks), but this skill rapidly decreased with training. This was due to the fact that CNNs at early stages of DT call too many events. As it trains further, it reduces the forecast rate, declaring fewer false positives at the expense of missing more events.

7.2 Transfer learning

Transfer learning (TL) has found broad application in atmospheric science, such as detecting gravity waves (González et al., 2022), improving extreme heatwave forecasts in climate models (Jacques-Dumas et al., 2022), subgrid-scale turbulence parameterization (Subel et al., 2021), image restoration (Guo et al., 2022) and parameter retrieval from raw dew point temperature profiles (Malmgren-Hansen et al., 2018).

TL involves pre-training a model on a larger dataset that is similar to the dataset of interest (source domain), then fine-tuning the model on the smaller target dataset (target domain). This approach is particularly beneficial when labeled data for the target task is limited, as it allows the model to exploit learned features and representations from the larger dataset to enhance its performance on the smaller dataset. With this strength, TL has shown its power in forecasting, combining the data from a climate model (Rasp & Thuerey, 2021) or a dynamical model (Mu et al., 2020) with the observational record to improve medium-range weather forecasting and ENSO prediction.

In this section, we apply TL to leverage our MM dataset to predict events in the reanalysis data. As a quasi-geostrophic model, MM has complexity between full climate models (e.g., Rasp & Thuerey, 2021) and low order models (e.g., Mu et al., 2020) used in previous transfer learning studies. The overall process is to first 'pre-train' a CNN on the MM model dataset, learning to capture the characteristic features of blocking. While significantly simplified, the MM model is skillful in representing atmospheric variability (Lucarini & Gritsun, 2020), but more importantly provides extensive positive and negative cases to learn from, supporting optimal CNN training, as demonstrated in Sec. 4. After pre-training, our CNN is then fine-tuned on the ERA5 dataset, where the weights are modified to account for biases in the MM model, and the parameter scales are calibrated.

In most applications of TL, only the weights in the last few layers of a neural network are fine-tuned on the target domain (Yosinski et al., 2014; Hussain et al., 2019; Talo et al., 2019). Following this convention, we only retrain the last layer of the CNN on ERA5 while keeping the other layers frozen. This allows the CNN to correct biases it inherits from MM, but not to fall back into the poorly constrained limit we reached with direct training. We also tried retraining other single layers, but retraining the last layer performed the best. To avoid overfitting, we set the learning rate to 1/10 the learning rate of pre-training.

We tested different lengths of pre-training and then evaluated the performance of the resulting models with the peak precision and recall in the transfer-learning phase. The results show that CNN parameters taken at earlier pre-training epochs show better peak performance after transfer learning (results not shown). This suggests that overfitting on the source domain cannot be fully corrected by fine-tuning on the target domain. For the displayed results in Figs. 8, 9 and 10, we use a pre-training of 2 epochs for $D = 5$, and 1 epoch for $D = 7$. Given the 1000k days of MM integration we had at our disposal, this means that the neural network has explored more than 70,000 unique nascent blocking states (all of them twice, for $D = 5$) before seeing any of the 273 events in ERA5.

We follow a similar procedure as with DT to assess the ensemble-average performance. We pre-train 10 CNNs with the 1000k-day MM dataset; the only differences are due to randomness in the initialization and training data shuffling. We then carry out a 10-fold cross-validation procedure with 90:10 splits: for each split, we perform TL fine-tuning on the 10 pre-trained CNNs. We compute the mean precision and recall for each split. The results in the TL columns of Fig. 8 show the mean and 2-standard deviation bounds across all the splits.

Compared to DT, TL begins with a higher precision but lower recall due to pre-training. With additional fine-tuning, the precision stays almost unchanged, while the recall grows markedly. The network is able to increase the number of events that it can capture (lowering the number of false negatives) with minimal degradation in reliability of its forecast (that is, only slightly increasing the false positive rate).

Uncertainty in the precision is dominated by differences in the true positive events between the splits; consequently, the 2-standard deviation error bounds are comparable for DT and TL. The recall is less sensitive to differences among the splits, however, and at least for the $D = 5$ case, there is noticeably less spread across the splits with transfer learning. This is understandable because recall, by definition, doesn't depend on the positive rate of the test dataset, which varies a lot for small data sets (around 27 states in each test set after splitting). On the other hand, precision relies on the positive rate of the test dataset, so it has more intrinsic variability.

We still evaluate the overall performance by Eq. (6). Focusing first on $D = 5$ events, the best mean performance with DT is a precision of 0.45 and recall of 0.61, which is realized at Epoch 3. With TL, we achieve an average performance with a similar precision of 0.45 and higher recall 0.82 (at Epoch 4). A noticeable advantage of TL is the significantly reduced variance in recall compared to DT, indicating TL's superior robustness in prediction, attributed to its enhanced capacity for capturing predictive features. For $D = 7$ day events, the best mean performance with DT is a precision of 0.21 and recall of 0.48, achieved after 3 epochs. TL, however, achieves a precision of 0.22 and recall of 0.76 at Epoch 6.

To ensure that these gains in recall are statistically significant, we conducted a Wilcoxon signed-rank test (Conover, 1999). Fig. 9 shows histograms of the difference in precision and recall between direct training and transfer learning. For example, each of the 10 values in the histogram for $D = 5$ is defined for a specific train-test split, evaluated by subtracting the mean precision (recall) of 10 randomly initialized TL models taken at Epoch 4 from the mean precision (recall) of 10 randomly initialized DT models taken at Epoch 3. The spread here stems primarily from the fluctuation in 10 small-size test sets, not uncertainty in the networks due to randomness in training. The values for small-size test sets are taken at the same epoch of the best mean performance.

The average recall with TL surpasses that of DT by 34% ($p = 0.001$) for 5 d events and by over 50% ($p = 0.002$) for 7 d events. While there is not a significant difference between the TL and DT precision, it is critical that transfer learning was able to improve the recall without sacrificing precision. One could easily inflate the recall by declaring more positive cases, but without any skill, the precision would suffer and approach the climatological rate.

7.3 What has transfer learning learned?

When we show ERA5 events to CNNs first trained on the MM dataset, what exactly is the CNN learning to improve the recall? For example, do the key geographical regions and levels (Fig. 5) retain the same level of significance? It is reasonable to expect that this might not be the case. In the MM dataset, the duration of the Atlantic blockings could be related to upstream flow, specifically to the structure of the wave train at the blocking onset. The mechanism for blocking in the real world is more complicated, and the correlated pattern may shift, intensify, and/or weaken. To address these questions, we compare the SHAP values of the pre-trained CNNs when directly applied to ERA5 (i.e., without fine-tuning) to the SHAP values of the CNN after 4 epochs of fine-tuning, as shown in row *a* and row *b* of Fig. 10. The most evident difference after fine-tuning is a decrease in the amplitude of the SHAP values. This is because the climatological rate of positive blocking events in ERA5 is higher: almost 1/3 of nascent blocked states persist for 5 d in ERA5, compared to about 1/5 in MM. As the expected fraction

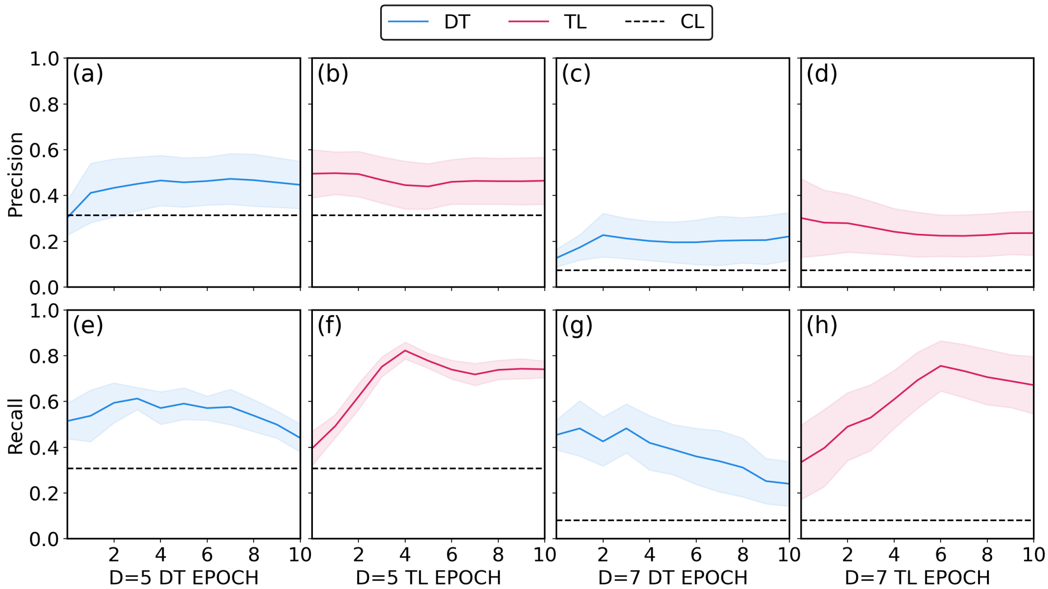

Figure 8. Comparison of CNN forecast skill between direct training (DT, blue) and transfer learning (TL, red). Panels (a,b) compare the precision of DT training epoch and of TL fine-tuning epoch for D=5 (standard blocking events). (e,f) compare the recall of DT training epoch and of TL fine-tuning epoch for D=5. (c,d) compare the same quantities as (a,b) for D=7. (g,h) compare the same quantities as (e,f) for D=7 (longer blocking events). The black dashed line indicates the climatological event rate p. The shading shows a two-standard deviation uncertainty bound, as detailed in the text.

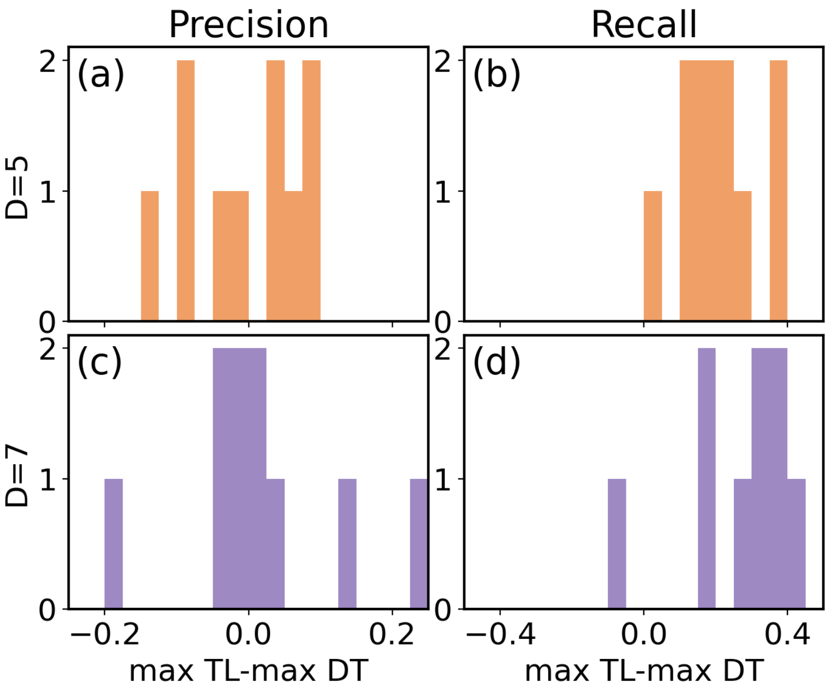

Figure 9. Histograms of the performance gap between the best performing CNNs obtained with transfer learning versus the best performing CNNs obtained with direct training, for precision and recall. (a) is the performance gap of precision for 5 d events. (b) is that of recall for 5 d events. (c) and (d) are of precision and recall for 7 d events. "Best performing" was determined by stopping the training procedure at the epoch when the best overall balance between high precision and recall was achieved in the mean (solid lines in Fig. 8). The 90:10 split yields 10 different CNN scores, and the differences between pairs of TL and DT based CNNs, scored on the same test split, are shown.

of events is larger, $\hat{q}(\mathbf{x}) - \mathbb{E}[\hat{q}(\mathbf{x})]$ from equation (7) will be smaller, and the SHAP value increments $\phi_i(\hat{q}, \mathbf{x})$ will tend to be smaller. It is the sum of the SHAP values that build up the probability for a $Y = 1$ prediction; for a more likely event, one does not need to build up the probability as much, so fine-tuning quickly adjusts the weights.

To assess the more subtle change in the relative contribution of each feature on the predicted result after transfer learning, we show the difference in the normalized composite map $\Delta\phi$ in row d of Fig. 10. $\Delta\phi$ is defined for each input i (i.e., geopotential height Z at a particular latitude, longitude, and pressure level) by $\Delta\phi_i \equiv \max\left(\frac{\overline{\phi}_i^{\mathrm{TL}}}{\frac{1}{d}\sum_{j=1}^{d}\overline{\phi}_j^{\mathrm{TL}}}, 0\right) - \max\left(\frac{\overline{\phi}_i}{\frac{1}{d}\sum_{j=1}^{d}\overline{\phi}_j}, 0\right)$. The maximum function is used to avoid spurious negative SHAP values, which should not arise in a composite of true positive events, as discussed in the context of Fig. 5. The normalization makes the total integral of the SHAP values the same for both cases, so that one can focus on where the CNN is using information, as opposed to the overall reduction of the SHAP values driven by the difference in rates.

The "normalized" SHAP values increase mainly in region 4 (the region right around the block), and additionally over Quebec and Atlantic Canada, a region less used for predictions with the MM model. The SHAP values decrease in a relative sense over regions 1 (Florida and the Gulf), 2 (North Atlantic Ocean), 3 (northeastern North America), and central North America. This change in relative importance reveals a general de-emphasis of the regions farther upstream and an increased emphasis on regions more immediately upstream. This indicates that while it is still upstream information that is most important for predicting a persistent blocking state in ERA5, the structure and westward extension of the wave train has changed.

For further insight, we compare the SHAP value patterns with a more traditional method for understanding predictability: composite analysis. Fig. 6 shows composite maps of nascent blocks that evolve into persistent events in the MM model and ERA5. Persistent blocks are associated with wave activity south and west of the blocking region in both the model and reanalysis, but the pattern shifts. The wave train in MM initially arcs westward before turning southward, with a strong center of high pressure east of Florida, while the wave train in ERA5 arcs more to southwest at first, then further westward.

The SHAP values change over Quebec, capturing this shift in the wave train, but overall the CNN seems to shift to more local information with transfer learning. We speculated that the dry, quasi-geostrophic MM model overemphasizes long-range teleconnections. It only captures deformation scale dynamics, and this only at low resolution, and so lacks smaller, local modes of instability, e.g., instability associated with latent heat release due to precipitation, present in our atmosphere. The CNN makes more use of these local features when predicting the persistence of blocks, but still focuses on the upstream flow, consistent with our intuition.

Finally, we contrast the feature importance analysis of the CNN with transfer learning (Fig. 10 row b) to that of the CNNs trained only directly on the ERA5 output (Fig. 10 row c). DT struggles to develop nuanced features with limited data. The SHAP values with DT are also more barotropic than those with TL. Moreover, in general, the SHAP values with TL capture finer details across a wider spatial range, while the SHAP values with DT are more localized. Geopotential height anomalies over Iceland, especially in the Z500 map, are more emphasized for TL than DT. The same applies to upstream anomalies over Florida and the Gulf of Mexico in the Z200 map. Additionally, the importance of geopotential height anomalies over the Atlantic, immediately upstream of the target region west of north Africa, is neglected in DT, though it appears in TL. This is closely correlated to the blocking event prediction from the ERA5 composite in Fig. 6, which does not show as strong composite Atlantic anomaly as in the MM model.

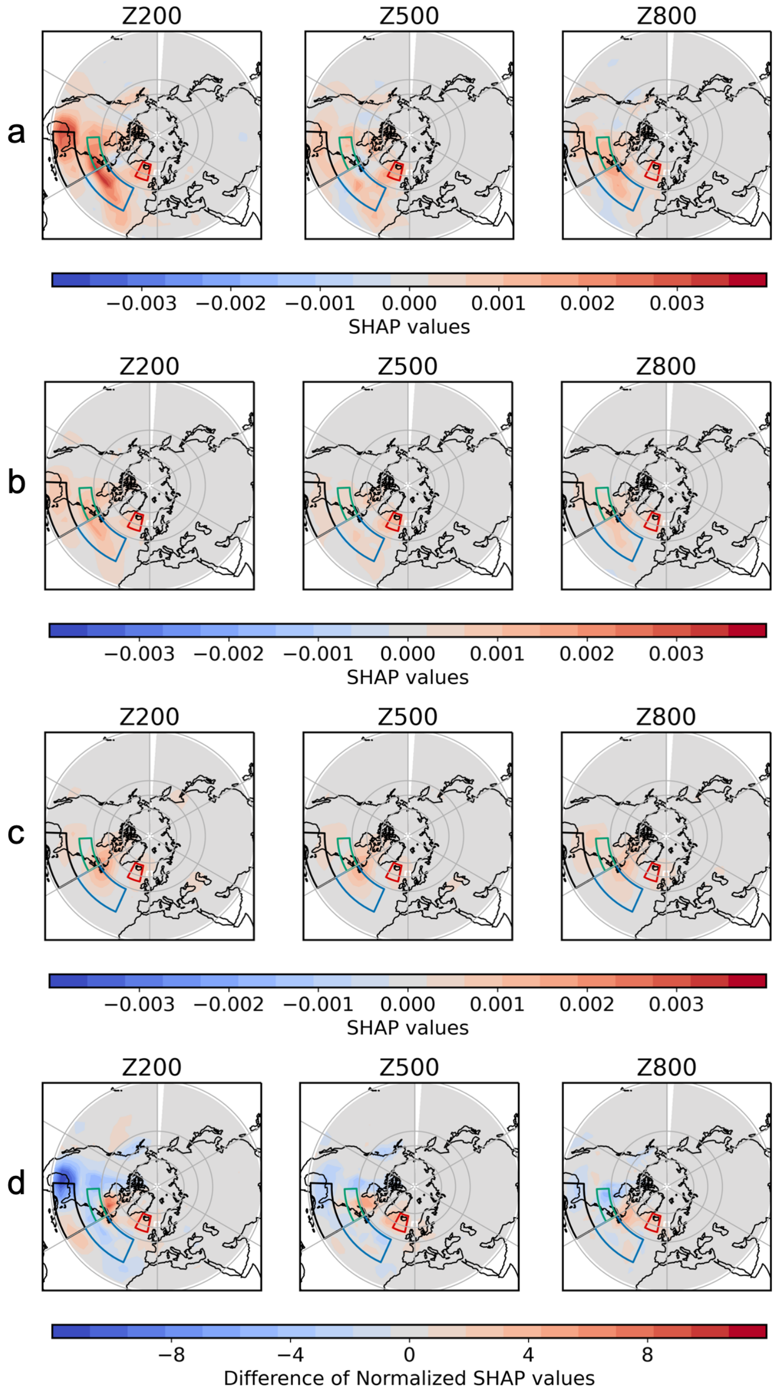

Figure 10. Rows 1 through 4 are composite maps of SHAP values, $\overline{\phi}$, for geopotential height (200, 500, and 800 hPa), averaged over true positive predictions of blocking events in ERA5 by the CNNs listed below. This is the same quantity shown in Fig.5, but now applied to ERA5 events. Row *a* shows $\overline{\phi}^{\mathrm{MM}}$ for the pre-trained CNNs before transfer learning (i.e., networks that have only learned from MM, but applied to ERA5). Row *b*: $\overline{\phi}^{\mathrm{TL}}$ of these pre-trained CNNs after fine-tuning. Row *c*: $\overline{\phi}^{\mathrm{DT}}$ of CNNs directly trained on ERA5 (i.e., networks that never saw the MM events). Row *d* shows the change in the SHAP values, $\Delta\phi$, between the first two rows, after normalization as detailed in the text. This quantifies the effect of transfer learning: positive values indicate that information from the region became more important for the prediction, while negative values indicate that anomalies in the region became less important for prediction.

In summary, the superiority of CNNs trained with transfer learning, as compared to direct training, appears to lie in their ability to leverage learned features from the pre-trained dataset, helping the network to take advantage of information further upstream of the blocking region. In either case the precision is modest: when the networks call an event, the rate of success is at best 50% higher than a naïve climatological forecast. Pre-training the network, however, has a significant impact on the recall, increasing the forecast rate to capture more events without decreasing the precision.

8 Conclusion

The impact of data-driven science on weather and climate science has grown substantially in recent years. In this paper, we suggest two data-driven approaches to help predict and understand atmospheric blocking events. First, given sufficient data, convolutional neural networks (CNNs) are capable of identifying subtle features that differentiate short-lived blocked states from those that persist for an extended period. Moreover, Explainable Artificial Intelligence methods, like SHAP feature importance analysis, can provide insight into what features matter most to this differentiation. Second, transfer learning has the potential to make data-driven forecasts possible for our atmosphere, making the most of the limited extreme events in the observational record by leveraging insight from longer, albeit imperfect, numerical simulations.

We began in a data-rich regime with the idealized Marshall-Molteni model, showing that a CNN can accurately predict the persistence of North Atlantic blocks in terms of both precision and recall. Leveraging SHAP feature importance analysis, we identified crucial regions for the prediction of persistent blocked states, given a nascent high-pressure anomaly. Our results suggest that incorporation of both local and non-local features is important for prediction skill.

To validate our discovery, we constructed a two-dimensional model that used only upstream anomalies over Florida and the Gulf of Mexico, and anomalies immediately upstream of the blocking region. The sparse model exhibited precision significantly above the climatological rate and recall nearly as good as the full CNN. It struggled, however, with false positives (and hence exhibited low precision relative to the CNN) which could not be improved within the log linear logistic regression framework. This suggests the CNN learns non-trivial relations in the upstream flow, extending all the way to the Pacific, to better discriminate between short-lived and long-lived blocks.

The challenge of conducting direct training on ERA5 data stems from the paucity of available events. Small training and test datasets make training and evaluation difficult. With the MM model, we observed a systematic degradation in forecast skill when the training data was limited, particularly for the recall statistic. Through transfer learning, we leverage the abundance of data generated by simplified dynamical models to enhance real-world forecasting. By pre-training a CNN on the MM model dataset and re-training the deepest layer on the ERA5 dataset, the recall was improved by 34% compared to a CNN developed with direct training alone for 5 d events, and over 50% for more extreme 7 d events, without any loss of precision.

In addition to advancing predictive skill, transfer learning in combination with SHAP analysis allowed us to compare the predictive features between weather systems in ERA5 and the idealized quasigeostrophic model. The bottom row of Fig. 6 reveals biases in the MM model, which appears overly dependent on upstream features over Florida and the Gulf of Mexico relative to blocks in ERA5. This approach provides a new angle of how a machine learning approach could guide the diagnosis and quantification of model biases. This said, the success of transfer learning results underscores the MM model's ability, despite its simplicity, to capture features that are important for predicting the persistence of blocked states in the real world. We believe that greater strides could be made by pre-training on a more advanced climate model, or even hindcasts in the subseasonal-to-seasonal (S2S) data set (Vitart et al., 2017; Finkel et al., 2023). We expect our results will help inform large-scale efforts to incorporate AI into operational forecasts, such as the AIFS model (Lang et al., 2024), which already employs transfer learning in a different form.

The methods presented here are not limited to the context of blocking events, and can be generalized to the study of other challenging natural phenomena, especially in scenarios where data may be limited, and the potential influencing factors are complex (e.g. heat domes (Li et al., 2024)). An immediate future goal is to push further on the physical and dynamical mechanisms that causes the differences in prediction mechanisms for ERA5 and MM model. Another goal is to adapt the present approach to investigate the statistical behavior and mechanisms for the onset of the blocking events.

Appendix A Marshall-Molteni Model

The Marshall-Molteni (MM) model state is specified by potential vorticity q_j in three layers of the atmosphere, $j = 1, 2, 3$, corresponding to pressure levels 200, 500, and 800 hPa. q_j evolves according to quasi-geostrophic dynamics as

$$\partial_t q_j + J(\psi_j, q_j) = -D_j + S_j \tag{A1}$$

where ψ_j is the streamfunction in layer j, related to q_j as

$$q_1 = \Delta\psi_1 - (\psi_1 - \psi_2)/R_1^2 + f \tag{A2}$$
$$q_2 = \Delta\psi_2 + (\psi_1 - \psi_2)/R_1^2 - (\psi_2 - \psi_3)/R_2^2 + f \tag{A3}$$
$$q_3 = \Delta\psi_3 + (\psi_2 - \psi_3)/R_2^2 + f(1 + h/H_0). \tag{A4}$$

Here, Δ is the horizontal Laplacian operator, $R_1 = 761$ km and $R_2 = 488$ km are the Rossby deformation radii in layers 1 and 2, $f = 2\Omega\cos\phi$ is the latitude-dependent Coriolis parameter, and h is the orography of the surface, rescaled by the constant H_0. The operator D_j combines all dissipative terms, including radiative damping, surface friction and hyper-diffusion to crudely parametrize small scale diffusion, but is also necessary for numerical stability:

$$\begin{aligned} -D_1 =& (\psi_1 - \psi_2)/(\tau_R R_1^2) - R^8\Delta^4 q_1/(\tau_H \lambda_{max}^4) \\ -D_2 =& -(\psi_1 - \psi_2)/(\tau_R R_1^2) + (\psi_2 - \psi_3)/(\tau_R R_2^2) - R^8\Delta^4 q_2'/(\tau_H \lambda_{max}^4) \\ -D_3 =& -(\psi_2 - \psi_3)/(\tau_R R_2^2) - EK_3 - R^8\Delta^4 q_3'/(\tau_H \lambda_{max}^4). \end{aligned} \tag{A5}$$

The forcing, S_j is computed from observed data to inject energy into the system and give the model a realistic mean state:

$$S_j = \overline{J(\psi_j, q_j)} + \overline{D}_j \tag{A6}$$

The data to construct S_j were drawn from the 1983–1992 winter (DJF) climatology of the ERA40 reanalysis provided by ECMWF.

Appendix B Acronyms and definitions

Here we list the important acronyms and definitions in this paper for the convenience of the readers.

- CNN: Convolutional Neural Network - A type of deep learning model particularly effective for analyzing visual data, using convolutional layers to automatically detect and learn patterns.

- SHAP: Shapley Additive ExPlanation - A method to explain the output of machine learning models by attributing contributions of individual features based on cooperative game theory.
- MM: Marshall-Molteni - Refers to the 3-layer QG model by (Marshall & Molteni, 1993) related to atmospheric dynamics, often used in the context of studying large-scale weather patterns and teleconnections.
- QG: Quasi-Geostrophic - A simplified model in geophysical fluid dynamics that describes large-scale atmospheric and oceanic flows, assuming a balance between pressure gradient and Coriolis forces.
- XAI: Explainable Artificial Intelligence - A subfield of AI focused on making the outputs and processes of machine learning models transparent and understandable to humans.
- DG: Dole & Gordon index (Dole & Gordon, 1983)- An index developed by Dole and Gordon to quantify atmospheric blocking events, which are large-scale pressure systems that can disrupt normal weather patterns.
- DT: Direct Training - A machine learning approach where a model is trained directly on a specific dataset without additional pre-training or transfer learning techniques.
- TL: Transfer Learning - A machine learning technique where a pre-trained model is adapted to a new but related task, leveraging the knowledge gained from the original task to improve performance.
- Z: Geopotential height.
- $Z_B(t)$: Anomalous geopotential height in our target blocking region in the North Atlantic, shown in Fig. 1.
- T: Number of consecutive days of a blocked state.
- M: Threshold of geopotential height anomaly in blocking events criteria.
- D: Threshold of consecutive days in blocking events criteria.
- $\boldsymbol{X}$: Full model state vector.
- Y: Indicator of whether a blocked state persisted.
- $q(\boldsymbol{x}(t))$: Conditional probability that a blocked state $\boldsymbol{x}(t)$ will persist.
- $L(q)$: Binary cross entropy loss function used for classification problem.
- Precision: $\frac{\text{True positives}}{\text{True positives+False positives}}$
- Recall: $\frac{\text{True positives}}{\text{True positives+False negatives}}$

Open Research Section

The data from the Marshall-Molteni model were generated using a Fortran code provided by Valerio Lucarini and Andrey Gritsun (Lucarini & Gritsun, 2020). The Fortran code, along with the Python code for computing SHAP values, transfer learning and producing plots is publicly available in the the open repository (Zhang, 2024). SHAP values were computed using the Python package DeepSHAP(Chen, 2022). The ERA5 reanalysis datasets from ECWMF were used for data preprocessing and ML model training and testing (Hersbach et al., 2020).

Acknowledgments

We thank Valerio Lucarini and Andrey Gritsun for sharing their Marshall-Molteni Fortran code. We also thank Pedram Hassanzadeh and the anonymous reviewers for many helpful comments and suggestions that strengthend the paper. This work was supported by the Army Research Office, grant number W911NF-22-2-0124. EPG acknowledges support from the National Science Foundation through award OAC-2004572. J. F. is supported through the MIT Climate Grand Challenge on Weather and Climate Extremes, and the Virtual Earth Systems Research Institute (VESRI) at Schmidt Sciences.

References

Barnes, E. A., & Hartmann, D. L. (2010). Dynamical feedbacks and the persistence of the nao. *Journal of the Atmospheric Sciences*, *67*(3), 851 - 865. Retrieved from `https://journals.ametsoc.org/view/journals/atsc/67/3/2009jas3193.1.xml` doi: 10.1175/2009JAS3193.1

Berckmans, J., Woollings, T., Demory, M.-E., Vidale, P.-L., & Roberts, M. (2013). Atmospheric blocking in a high resolution climate model: influences of mean state, orography and eddy forcing. *Atmospheric Science Letters*, *14*(1), 34–40.

Chan, P.-W., Hassanzadeh, P., & Kuang, Z. (2019). Evaluating Indices of Blocking Anticyclones in Terms of Their Linear Relations With Surface Hot Extremes. *Geophysical Research Letters*, *46*(9), 4904-4912. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019GL083307` doi: https://doi.org/10.1029/2019GL083307

Charney, J. G., & DeVore, J. G. (1979). Multiple flow equilibria in the atmosphere and blocking. *Journal of Atmospheric Sciences*, *36*(7), 1205–1216.

Chen, H. (2022, May). *suinleelab/deepshap: Nature communications code [Software]*. Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.6585445` doi: 10.5281/zenodo.6585445

Conover, W. J. (1999). *Practical nonparametric statistics* (Vol. 350). john wiley & sons.

Davini, P., & D'Andrea, F. (2020). From CMIP3 to CMIP6: Northern Hemisphere Atmospheric Blocking Simulation in Present and Future Climate. *Journal of Climate*, *33*(23), 10021 - 10038. Retrieved from `https://journals.ametsoc.org/view/journals/clim/33/23/jcliD190862.xml` doi: https://doi.org/10.1175/JCLI-D-19-0862.1

Davini, P., & D'Andrea, F. (2016). Northern Hemisphere atmospheric blocking representation in global climate models: twenty years of improvements? *Journal of Climate*, *29*(24), 8823–8840.

Davini, P., Weisheimer, A., Balmaseda, M., Johnson, S. J., Molteni, F., Roberts, C. D., ... Stockdale, T. N. (2021). The representation of winter northern hemisphere atmospheric blocking in ecmwf seasonal prediction systems. *Quarterly Journal of the Royal Meteorological Society*, *147*(735), 1344–1363.

Dikshit, A., & Pradhan, B. (2021). Explainable AI in drought forecasting. *Machine Learning with Applications*, *6*, 100192. Retrieved from `https://www.sciencedirect.com/science/article/pii/S2666827021000967` doi: https://doi.org/10.1016/j.mlwa.2021.100192

Dole, R. M., & Gordon, N. D. (1983). Persistent anomalies of the extratropical Northern Hemisphere wintertime circulation: Geographical distribution and regional persistence characteristics. *Monthly Weather Review*, *111*(8), 1567–1586.

d'Andrea, F., Tibaldi, S., Blackburn, M., Boer, G., Déqué, M., Dix, M., ... others (1998). Northern Hemisphere atmospheric blocking as simulated by 15 atmospheric general circulation models in the period 1979–1988. *Climate Dynamics*, *14*, 385–407.

Evans, K. J., & Black, R. X. (2003). Piecewise tendency diagnosis of weather regime transitions. *Journal of the Atmospheric Sciences*, *60*(16), 1941 - 1959. Retrieved from `https://journals.ametsoc.org/view/journals/atsc/60/16/1520-0469_2003_060_1941_ptdowr_2.0.co_2.xml` doi: 10.1175/1520-0469(2003)060⟨1941:PTDOWR⟩2.0.CO;2

Ferranti, L., Corti, S., & Janousek, M. (2015). Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. *Quarterly Journal of the Royal Meteorological Society*, *141*(688), 916-924. Retrieved from `https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2411` doi: https://doi.org/10.1002/qj.2411

Finkel, J., Webber, R. J., Gerber, E. P., Abbot, D. S., & Weare, J. (2021). Learning Forecasts of Rare Stratospheric Transitions from Short Simulations. *Monthly Weather Review*, *149*(11), 3647 - 3669. Retrieved from https://journals.ametsoc.org/view/journals/mwre/149/11/MWR-D-21-0024.1.xml doi: https://doi.org/10.1175/MWR-D-21-0024.1

Finkel, J., Webber, R. J., Gerber, E. P., Abbot, D. S., & Weare, J. (2023). Data-Driven Transition Path Analysis Yields a Statistical Understanding of Sudden Stratospheric Warming Events in an Idealized Model. *Journal of the Atmospheric Sciences*, *80*(2), 519 - 534. Retrieved from https://journals.ametsoc.org/view/journals/atsc/80/2/JAS-D-21-0213.1.xml doi: https://doi.org/10.1175/JAS-D-21-0213.1

González, J. L., Chapman, T., Chen, K., Nguyen, H., Chambers, L., Mostafa, S. A., ... Yue, J. (2022). Atmospheric Gravity Wave Detection Using Transfer Learning Techniques. In *2022 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)* (p. 128-137). doi: 10.1109/BDCAT56447.2022.00023

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press. (http://www.deeplearningbook.org)

Guo, Y., Wu, X., Qing, C., Su, C., Yang, Q., & Wang, Z. (2022). Blind Restoration of Images Distorted by Atmospheric Turbulence Based on Deep Transfer Learning. *Photonics*, *9*(8). Retrieved from https://www.mdpi.com/2304-6732/9/8/582 doi: 10.3390/photonics9080582

Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019, Sep 01). Deep learning for multi-year ENSO forecasts. *Nature*, *573*(7775), 568-572. Retrieved from https://doi.org/10.1038/s41586-019-1559-7 doi: 10.1038/s41586-019-1559-7

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... Thépaut, J.-N. (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999-2049. Retrieved from https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803 doi: https://doi.org/10.1002/qj.3803

Hoskins, B. J., James, I. N., & White, G. H. (1983, July). The Shape, Propagation and Mean-Flow Interaction of Large-Scale Weather Systems. *Journal of Atmospheric Sciences*, *40*(7), 1595-1612. doi: 10.1175/1520-0469(1983)040⟨1595:TSPAMF⟩2.0.CO;2

Hussain, M., Bird, J. J., & Faria, D. R. (2019). A study on cnn transfer learning for image classification. In *Advances in Computational Intelligence Systems: Contributions Presented at the 18th UK Workshop on Computational Intelligence, September 5-7, 2018, Nottingham, UK* (pp. 191–202).

Jacques-Dumas, V., Ragone, F., Borgnat, P., Abry, P., & Bouchet, F. (2022). Deep learning-based extreme heatwave forecast. *Frontiers in Climate*, *4*.

Jacques-Dumas, V., van Westen, R. M., Bouchet, F., & Dijkstra, H. A. (2023). Data-driven methods to estimate the committor function in conceptual ocean models. *Nonlinear Processes in Geophysics*, *30*(2), 195–216. Retrieved from https://npg.copernicus.org/articles/30/195/2023/ doi: 10.5194/npg-30-195-2023

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 1–54.

Kautz, L.-A., Martius, O., Pfahl, S., Pinto, J. G., Ramos, A. M., Sousa, P. M., & Woollings, T. (2022). Atmospheric blocking and weather extremes over the Euro-Atlantic sector – a review. *Weather and Climate Dynamics*, *3*(1), 305–336. Retrieved from https://wcd.copernicus.org/articles/3/305/2022/ doi: 10.5194/wcd-3-305-2022

Labe, Z. M., & Barnes, E. A. (2021). Detecting Climate Signals Using Explainable AI With Single-Forcing Large Ensembles. *Journal of Advances in Modeling Earth Systems*, *13*(6), e2021MS002464. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/

2021MS002464 (e2021MS002464 2021MS002464) doi: https://doi.org/10.1029/2021MS002464

Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., … Rabier, F. (2024). *Aifs - ecmwf's data-driven forecasting system.* Retrieved from https://arxiv.org/abs/2406.01465

Li, X., Mann, M. E., Wehner, M. F., Rahmstorf, S., Petri, S., Christiansen, S., & Carrillo, J. (2024). Role of atmospheric resonance and land–atmosphere feedbacks as a precursor to the june 2021 pacific northwest heat dome event. *Proceedings of the National Academy of Sciences*, *121*(4), e2315330121. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.2315330121 doi: 10.1073/pnas.2315330121

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18.

Lipovetsky, S., & Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, *17*(4), 319-330. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.446 doi: https://doi.org/10.1002/asmb.446

Liu, Y., Racah, E., Prabhat, Correa, J., Khosrowshahi, A., Lavers, D., … Collins, W. (2016). *Application of Deep Convolutional Neural Networks for Detecting Extreme Weather in Climate Datasets.*

Lucarini, V., & Gritsun, A. (2020). A new mathematical framework for atmospheric blocking events. *Climate Dynamics*, *54*(1-2), 575–598.

Lucente, D., Herbert, C., & Bouchet, F. (2022). Committor functions for climate phenomena at the predictability margin: The example of el niño southern oscillation in the jin and timmermann model. *Journal of the Atmospheric Sciences*. Retrieved from https://journals.ametsoc.org/view/journals/atsc/aop/JAS-D-22-0038.1/JAS-D-22-0038.1.xml doi: 10.1175/JAS-D-22-0038.1

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Lupo, A. R. (2021). Atmospheric blocking events: a review. *Annals of the New York Academy of Sciences*, *1504*(1), 5-24. Retrieved from https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/nyas.14557 doi: https://doi.org/10.1111/nyas.14557

Lupo, A. R., Mokhov, I. I., Akperov, M. G., Chernokulsky, A. V., Athar, H., et al. (2012). A dynamic analysis of the role of the planetary-and synoptic-scale in the summer of 2010 blocking episodes over the European part of Russia. *Advances in Meteorology*, *2012*.

Malmgren-Hansen, D., Nielsen, A. A., Laparra, V., & Valls, G. C. (2018). Transfer Learning with Convolutional Networks for Atmospheric Parameter Retrieval. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium* (p. 2111-2114). doi: 10.1109/IGARSS.2018.8518097

Marshall, J., & Molteni, F. (1993). Toward a dynamical understanding of planetary-scale flow regimes. *Journal of the atmospheric sciences*, *50*(12), 1792–1818.

Matsueda, M. (2009). Blocking Predictability in Operational Medium-Range Ensemble Forecasts. *SOLA*, *5*, 113-116. doi: 10.2151/sola.2009-029

McWilliams, J. C. (1980). An application of equivalent modons to atmospheric blocking. *Dynamics of Atmospheres and Oceans*, *5*(1), 43-66. Retrieved from https://www.sciencedirect.com/science/article/pii/037702658090010X doi: https://doi.org/10.1016/0377-0265(80)90010-X

Michelangeli, P.-A., & Vautard, R. (1998). The dynamics of Euro-Atlantic blocking onsets. *Quarterly Journal of the Royal Meteorological Society*, *124*(548), 1045–1070.

Miloshevich, G., Cozian, B., Abry, P., Borgnat, P., & Bouchet, F. (2023, Apr). Probabilistic forecasts of extreme heatwaves using convolutional neural networks in a regime of lack of data. *Phys. Rev. Fluids*, *8*, 040501. Retrieved from `https://link.aps.org/doi/10.1103/PhysRevFluids.8.040501` doi: 10.1103/PhysRevFluids.8.040501

Mu, B., Ma, S., Yuan, S., & Xu, H. (2020). Applying convolutional lstm network to predict el niño events: Transfer learning from the data of dynamical model and observation. In *2020 ieee 10th international conference on electronics information and emergency communication (iceiec)* (p. 215-219). doi: 10.1109/ICEIEC49280.2020.9152317

Mullen, S. L. (1987). Transient eddy forcing of blocking flows. *Journal of the Atmospheric Sciences*, *44*(1), 3–22.

Pelly, J. L., & Hoskins, B. J. (2003). A new perspective on blocking. *Journal of the atmospheric sciences*, *60*(5), 743–755.

Rampal, N., Gibson, P. B., Sood, A., Stuart, S., Fauchereau, N. C., Brandolino, C., ... Meyers, T. (2022). High-resolution downscaling with interpretable deep learning: Rainfall extremes over New Zealand. *Weather and Climate Extremes*, *38*, 100525. Retrieved from `https://www.sciencedirect.com/science/article/pii/S2212094722001049` doi: https://doi.org/10.1016/j.wace.2022.100525

Rasp, S., & Thuerey, N. (2021). Data-Driven Medium-Range Weather Prediction With a Resnet Pretrained on Climate Simulations: A New Model for WeatherBench. *Journal of Advances in Modeling Earth Systems*, *13*(2), e2020MS002405. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002405` (e2020MS002405 2020MS002405) doi: https://doi.org/10.1029/2020MS002405

Rex, D. F. (1950). Blocking Action in the Middle Troposphere and its Effect upon Regional Climate. *Tellus*, *2*(3), 196-211. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2153-3490.1950.tb00331.x` doi: https://doi.org/10.1111/j.2153-3490.1950.tb00331.x

Rudy, S. H., & Sapsis, T. P. (2023). Output-weighted and relative entropy loss functions for deep learning precursors of extreme events. *Physica D: Nonlinear Phenomena*, *443*, 133570.

Sasaki, Y., et al. (2007). The truth of the f-measure. 2007. *URL: https://www. cs. odu. edu/mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07. pdf [accessed 2021-05-26]*, *49*.

Scaife, A. A., Woollings, T., Knight, J., Martin, G., & Hinton, T. (2010). Atmospheric blocking and mean biases in climate models. *Journal of Climate*, *23*(23), 6143–6152.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *International conference on machine learning* (pp. 3145–3153).

Shutts, G. (1983). The propagation of eddies in diffluent jetstreams: Eddy vorticity forcing of 'blocking'flow fields. *Quarterly Journal of the Royal Meteorological Society*, *109*(462), 737–761.

Silva, S. J., Keller, C. A., & Hardin, J. (2022). Using an explainable machine learning approach to characterize Earth System model errors: Application of SHAP analysis to modeling lightning flash occurrence. *Journal of Advances in Modeling Earth Systems*, *14*(4), e2021MS002881.

Subel, A., Chattopadhyay, A., Guan, Y., & Hassanzadeh, P. (2021). Data-driven subgrid-scale modeling of forced Burgers turbulence using deep learning with generalization to higher Reynolds numbers via transfer learning. *Physics of Fluids*, *33*(3).

Talo, M., Baloglu, U. B., Yıldırım, Ö., & Acharya, U. R. (2019). Application of deep transfer learning for automated brain abnormality classification using MR images. *Cognitive Systems Research*, *54*, 176–188.

Tantet, A., van der Burgt, F. R., & Dijkstra, H. A. (2015). An early warning indicator for atmospheric blocking events using transfer operators. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *25*(3), 036406. Retrieved from https://doi.org/10.1063/1.4908174 doi: 10.1063/1.4908174

Tibaldi, S., & Molteni, F. (1990). On the operational predictability of blocking. *Tellus A*, *42*(3), 343–365.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., ... Zhang, L. (2017). The Subseasonal to Seasonal (S2S) Prediction Project Database. *Bulletin of the American Meteorological Society*, *98*(1), 163 - 173. Retrieved from https://journals.ametsoc.org/view/journals/bams/98/1/bams-d-16-0017.1.xml doi: 10.1175/BAMS-D-16-0017.1

Woollings, T., Barriopedro, D., Methven, J., Son, S.-W., Martius, O., Harvey, B., ... Seneviratne, S. (2018, Sep 01). Blocking and its Response to Climate Change. *Current Climate Change Reports*.

Yang, M., Luo, D., Li, C., Yao, Y., Li, X., & Chen, X. (2021). Influence of Atmospheric Blocking on Storm Track Activity Over the North Pacific During Boreal Winter. *Geophysical Research Letters*, *48*(17), e2021GL093863. Retrieved 2023-08-03, from https://onlinelibrary.wiley.com/doi/abs/10.1029/2021GL093863 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2021GL093863) doi: 10.1029/2021GL093863

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in neural information processing systems*, *27*.

Zappa, G., Masato, G., Shaffrey, L., Woollings, T., & Hodges, K. (2014a). Linking Northern Hemisphere blocking and storm track biases in the CMIP5 climate models. *Geophysical Research Letters*, *41*(1), 135–139.

Zappa, G., Masato, G., Shaffrey, L., Woollings, T., & Hodges, K. (2014b). Linking Northern Hemisphere blocking and storm track biases in the CMIP5 climate models. *Geophysical Research Letters*, *41*(1), 135–139. Retrieved 2023-08-02, from https://onlinelibrary.wiley.com/doi/abs/10.1002/2013GL058480 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/2013GL058480) doi: 10.1002/2013GL058480

Zhang, H. (2024, September). *hzhang-math/blockingshaptl: Code_blocking.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.13829703 doi: 10.5281/zenodo.13829703