

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

4 **Huan Zhang¹, Justin Finkel², Dorian S. Abbot³, Edwin P. Gerber¹, and**
5 **Jonathan Weare¹**

6 ¹Courant Institute of Mathematical Sciences, New York University

7 ²Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology

8 ³Department of the Geophysical Sciences, University of Chicago

9 **Key Points:**

- 10 • Given sufficient training data, convolutional neural networks can predict the main-
11 tenance of Atlantic blocking from an initial blocked state.
- 12 • Transfer learning from an idealized model to reanalysis data enables predictive skill
13 in the low data regime of the observational record.
- 14 • Feature importance analysis reveals the influence of upstream flow on blocking per-
15 sistence 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 quasigeostrophic 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 (Woollings et al., 2018; Matsueda, 2009; Ferranti et al., 2015). 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

66 the choice of blocking index itself (Tibaldi & Molteni, 1990; Dole & Gordon, 1983; Pelly
67 & Hoskins, 2003).

68 Two commonly used blocking indices (Tibaldi & Molteni, 1990; Dole & Gordon,
69 1983) highlight two essential features of a blocking *event*: (i) a large positive anomaly
70 of geopotential height that displaces the midlatitude jet, “blocking” the flow, that (ii)
71 persists for longer than typical synoptic variability. Often a 5 day threshold is invoked,
72 but the longer the flow remains in a blocked state, the more severe the implications, ei-
73 ther for extended cold/hot conditions or an increased likelihood of compound storm events
74 (e.g., back-to-back storms, which can dramatically increase the potential for damage; Kautz
75 et al., 2022). The persistence of blocking is the focus of our study: given the onset of a
76 blocked state, what is the likelihood that the flow will remain blocked for an extended
77 period, 5 days for a standard event, or up to 9 days for more extreme cases? We take
78 a data-driven approach, training a convolutional neural network to identify persistent
79 blocks at the onset of a blocked state.

80 To understand blocking, various low-order models have been formulated to iden-
81 tify essential features. In an influential early work, Charney and DeVore (1979) mod-
82 eled blocking as one of two equilibrium states of a set of dynamical equations for a highly
83 truncated barotropic channel model. Others used low-order models to propose that the
84 positive feedback of synoptic-scale eddies on the blocking structure contributes to the
85 long-time maintenance of blocks (Hoskins et al., 1983; Shutts, 1983; McWilliams, 1980).
86 While these low-order models have provided useful physical insight, realistic land-sea in-
87 teractions, topography, and other factors present in the real world limit their applica-
88 tion. Comprehensive models, on the other hand, are becoming skillful in simulating re-
89 alistic blocking, but their complexity makes it challenging to isolate the essential mech-
90 anism(s), and expensive to simulate numerous events.

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

101 The MM model allows us the freedom to develop and test methods in a data-rich
102 setting. How well can a data-driven method identify persistent events as a function of
103 the input data you allow it? Following work by Labe and Barnes (2021) and Rampal et
104 al. (2022), can so-called Explainable Artificial Intelligence (XAI) techniques provide phys-
105 ical insight into both the AI methods and the model itself? We show that Shapley Ad-
106 ditive ExPlanation (SHAP) analysis reveals key regions upstream of the blocking center
107 that enable prediction, and use this to construct low-order models the can be inter-
108 preted in the context of prior work.

109 Our ultimate goal, however, is to forecast and understand the maintenance of blocks
110 in our atmosphere, for which we shift the focus to ERA5 reanalysis (Hersbach et al., 2020).
111 For the most extreme case of a 9-day block in the North Atlantic, only 18 have occurred
112 in the historical record (See Tab. 3). What chance does a data-driven approach have?
113 To address the problem of limited data, we apply transfer learning: first we train a con-
114 volutional neural network on the MM model to learn the basic features of blocking, and
115 then we re-train it on the limited ERA5 data to calibrate it for the real atmosphere. We
116 find that pre-training on the MM model yields a better predictor than when we train
117 the same network on ERA5 alone, proving the efficacy of the transfer learning approach.

118 The remainder of this paper is organized as follows. Section 2 introduces the Marshall-
 119 Molteni (MM) model. Sections 3 and 4 define our choice of blocking index and block-
 120 ing event criteria, and formulate an objective function for machine learning. Section 5
 121 discusses our convolutional neural network structure and training details. We first fo-
 122 cus exclusively on the MM model in sections 6 and 7, applying XAI techniques to vi-
 123 sualize the important features for prediction and testing the results by building a sparse
 124 model with features guided by the XAI. We also suggest physical interpretations for these
 125 predictive features. Finally, we turn to the ERA5 data set in Section 8, applying trans-
 126 fer learning to improve the prediction of persistent blocks in ERA5, especially for more
 127 extreme events. SHAP analysis shows how transfer learning has modified the CNN to
 128 adapt to the new data set, but preserves the use of key upstream regions for prediction.

129 2 Marshall-Molteni Model

Marshall and Molteni (1993) developed a 3-layer model of the atmosphere to study
 atmospheric low-frequency variability. We use a Northern Hemisphere only version of
 the model developed by Lucarini and Gritsun (2020) with 6210 degrees of freedom. We
 refer the reader to that paper for a complete description, but review key details here. The
 Marshall-Molteni (MM) model state is specified by potential vorticity q_j in three lay-
 ers 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

$$\partial_t q_j + J(\psi_j, q_j) = -D_j + S_j \quad (1)$$

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

$$q_1 = \Delta\psi_1 - (\psi_1 - \psi_2)/R_1^2 + f \quad (2)$$

$$q_2 = \Delta\psi_2 + (\psi_1 - \psi_2)/R_1^2 - (\psi_2 - \psi_3)/R_2^2 + f \quad (3)$$

$$q_3 = \Delta\psi_3 + (\psi_2 - \psi_3)/R_2^2 + f(1 + h/H_0). \quad (4)$$

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 Cori-
 olis 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:

$$\begin{aligned} -D_1 &= (\psi_1 - \psi_2)/(\tau_R R_1^2) - R^8 \Delta^4 q_1 / (\tau_H \lambda_{max}^4) \\ -D_2 &= -(\psi_1 - \psi_2)/(\tau_R R_1^2) + (\psi_2 - \psi_3)/(\tau_R R_2^2) - R^8 \Delta^4 q_2' / (\tau_H \lambda_{max}^4) \\ -D_3 &= -(\psi_2 - \psi_3)/(\tau_R R_2^2) - EK_3 - R^8 \Delta^4 q_3' / (\tau_H \lambda_{max}^4). \end{aligned} \quad (5)$$

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

$$S_j = \overline{J(\psi_j, q_j)} + \overline{D}_j \quad (6)$$

130 The data to construct S_j were drawn from the 1983–1992 winter (DJF) climatol-
 131 ogy of the ERA40 reanalysis provided by ECMWF. The model is run with T31 horizon-
 132 tal resolution (corresponding to 90 longitude \times 23 latitude gridpoints across the north-
 133 ern hemisphere). All model output fields, as well as the reanalysis used later, are aver-
 134 aged daily. The climatology of Marshall-Molteni model is shown in the supplemental ma-
 135 terials, and we compare its blocking statistics with ERA5 reanalysis in the next section.

136 3 Blocking index

137 In this study, we use the ‘‘DG’’ index (Dole & Gordon, 1983) to define blocking events.
 138 This is an anomaly-based blocking index, but has been shown to capture the same es-
 139 sential features of blocking as other indices, e.g., that of Tibaldi and Molteni (1990).

140 We compute this index by transforming the spherical harmonic representation of
 141 ψ into approximate geopotential height, Z , on a Gaussian grid for latitude and a uni-
 142 form grid for longitude. The approximation is the choice of a fixed Coriolis parameter
 143 f_0 to convert from ψ to Z , which leads to minimal distortion over our midlatitude area
 144 of focus. A blocking event is said to occur at a specific location when Z stays above a
 145 tunable geopotential height anomaly threshold, M , for at least 5 consecutive days. In
 146 their paper, Dole and Gordon (1983) tested statistics for varying M values, ranging from
 147 50 m to 250 m, with subsequent studies adopting different thresholds (Chan et al. (2019),
 148 Tab. 2). For our investigation, we calibrated $M = 100$ m for our MM model simula-
 149 tion to roughly match the blocking fraction computed from ERA5 reanalysis data, where
 150 we used the threshold $M = 150$ m as in Mullen (1987).

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

158 4 Probabilistic forecasting and event definition

159 We aim to study the *maintenance* of blocks rather than their *onset*. Precisely, we
 160 formulate the question as the classification problem posed in Fig. 2: given a nascent blocked
 161 state, i.e., the state on a day that geopotential height anomalies over the North Atlantic
 162 first exceed the threshold M , can we immediately predict whether the flow will remain
 163 blocked for 5 or more days – evolving into a blocking *event* – or will the flow return back
 164 towards the climatological state before 5 days have passed? In the MM model, nascent
 165 blocked states evolve into 5-day persistent blocking events approximately 1/5th (21%)
 166 of the time on average, more often fading back towards climatology. Given only the state
 167 at the time of blocking onset, can a data-driven method accurately identify the rarer cases
 168 that will persist for more than 5 consecutive days?

To formulate this classification problem mathematically, we denote the full model
 state by \mathbf{X} and further introduce a variable T for the running duration of a blocked state:

$$T = \{\text{days since } Z_B < M\}. \quad (7)$$

Note that $Z_B(t)$ is determined by the state vector $\mathbf{X}(t)$ at any time t , but $T(t)$ retains
 some memory of previous states and thus is not fully determined by $\mathbf{X}(t)$. For exam-
 ple, 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 “ $\mathbf{X}(t)$ is the begin-
 ning of a blocking event” if

$$T(t) = 1 \quad \text{and} \quad T(t + D - 1) = D. \quad (8)$$

169 The condition $T(t+D-1) = D$ only holds when there are at least D consecutive days
 170 with $Z_B(t) \geq M$ starting from t . We can see an example of this in Fig. 2 at day 24,
 171 for both a block of duration 5 and 7 days. Here, $T(24) = 1$, and $T(28) = 5$, trigger-
 172 ing the condition for $D = 5$. The flow remains blocked through $T(30) = 7$, such that
 173 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:

$$q(\mathbf{x}(t)) = \mathbb{P}[T(t + D - 1) = D \mid \mathbf{X}(t) = \mathbf{x}(t), T(t) = 1]. \quad (9)$$

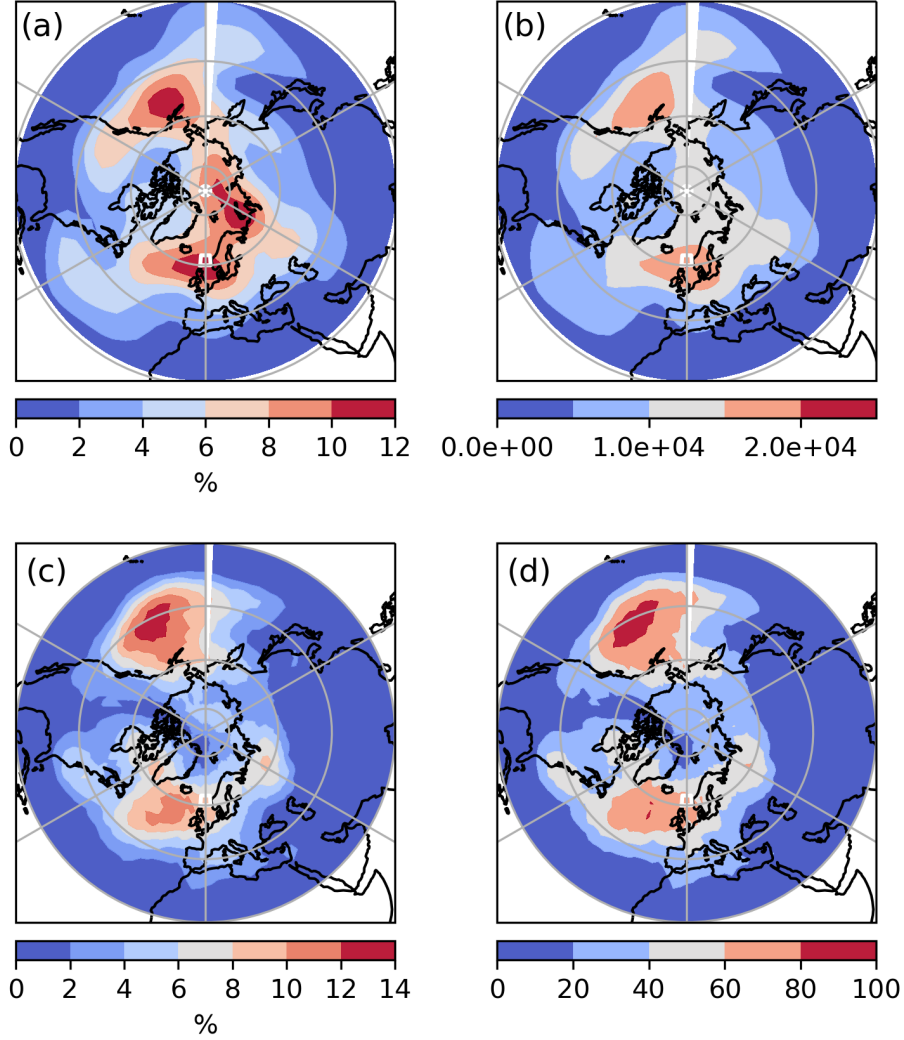


Figure 1. (a) blocking fraction (the percent of days with $T \geq 5$ days) for MM model data with $M = 100$ m. (b) total blocking event counts for MM model data during the simulation. (c) blocking fraction for ERA5 reanalysis data with $M = 150$. (d) total blocking event for ERA5 reanalysis data with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

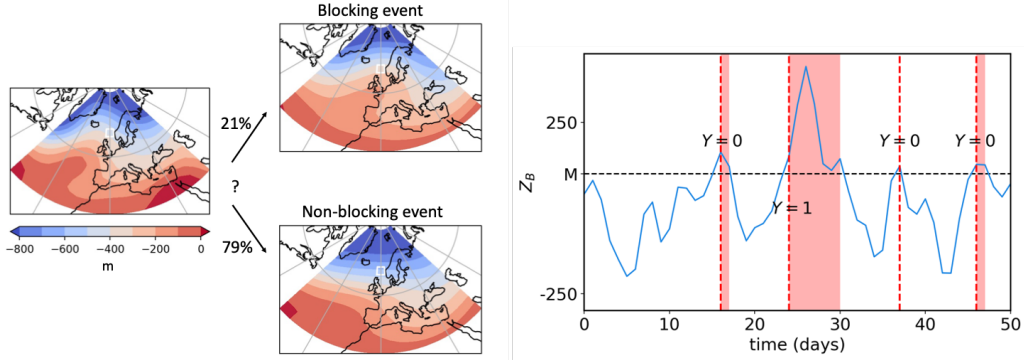


Figure 2. Left: The blocking persistence problem: given a nascent blocked state, the goal is to forecast whether it will persist into a long-lasting blocking event, or quickly return to climatology. The percentile represents the climatological probability. Right: A sample trajectory of $Z_B(t)$, the anomaly of geopotential height defined in Sec. 3. 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.

174 Unless otherwise specified, we adopt $D = 5$ to maintain consistency with the common
 175 blocking indices (Tibaldi & Molteni, 1990; Dole & Gordon, 1983; Pelly & Hoskins, 2003).
 176 We also consider more extreme events with $D = 7$ and $D = 9$.

177 5 Convolutional Neural Network Training and Performance

178 Convolutional Neural Networks (CNN) have gained widespread application in prob-
 179 abilistic forecasting problems (Miloshevich et al., 2023; Ham et al., 2019; Liu et al., 2016)
 180 for their outstanding performance on multidimensional data sets with spatial structure.
 181 A CNN differs from a dense neural network in the use of convolutional layers with shared
 182 weights and biases across layers within the network, designed to extract features that
 183 exhibit translation invariance across the input space (Goodfellow et al., 2016). Originally
 184 developed in the context of image processing, CNN excels in scenarios where target fea-
 185 tures, such as the face of a cat, may appear at different places within the training im-
 186 age. Convolutional layers allow the network to efficiently learn these features, combin-
 187 ing information across multiple images. In our context, atmospheric eddies and Rossby
 188 waves share similar dynamics across all longitudes. A CNN can potentially more effec-
 189 tively extract these dynamics, while still learning how they vary with longitude and zonal
 190 asymmetries induced by topography, etc.

191 The structure of the CNN in this investigation follows Miloshevich et al. (2023) and
 192 is shown in Fig. 3. It consists of a three-layer architecture, combining convolutional fil-
 193 ters followed by ReLU activations. Specifically, we use 32 and 64 filters (3×3) for the
 194 first and last two convolutional layers. Between each pair of convolutional layers is a max-
 195 pooling layer. The output is then flattened and passed to a dense layer with 64 neurons
 196 that produces 2 outputs. The output is then passed through a softmax function to form
 197 two normalized probabilities that sum to 1.

198 We performed experiments with alternative CNN structures and found that reduc-
 199 ing the widths of layers can mitigate overfitting, but this also reduces the performance

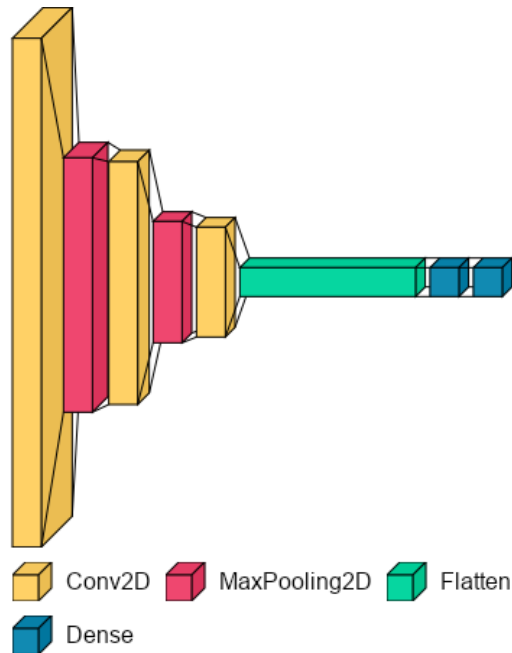


Figure 3. Convolutional Neural Network structure. The three convolutional layers (yellow) respectively use 32, 64 and 64 filters (3×3), followed by ReLu activations. Between each pair of convolutional layers is a max-pooling layer (red) with window size 2×2 . Then the output is flattened (green) and passed to a dense layer with 64 neurons that produces 2 outputs (blue). The output is then passed through a softmax function (blue).

200 at the best epoch (not shown). Therefore we adopt the architecture in Fig. 3 and use
 201 early-stopping to avoid overfitting, as detailed below.

202 5.1 Training and Test Datasets

203 To study whether a nascent blocked state will persist, we create a training and test
 204 set of all states where the flow has just become blocked: $\{(\mathbf{X}, T) | T = 1\}$, where \mathbf{X} are
 205 $18 \times 90 \times 3$ (latitudes \times longitudes \times pressure at levels of 200 hPa, 500 hPa, 800 hPa)
 206 grid maps of geopotential height from 20°N to 87°N . Our goal is to classify which of these
 207 cases persist into blocking events ($Y = 1$) versus states that do not ($Y = 0$). Fig. 2
 208 shows a sample time series with 4 instances of a nascent blocked state, $t = 16, 24, 38$
 209 and 47, only the second of which evolves into a persistent blocking event, $Y = 0, 1, 0,$
 210 and 0, respectively. For each case, the model must classify $Y = 0$ or $Y = 1$ given only
 211 \mathbf{X} at the onset time.

212 We examined the sensitivity of CNN model performance with respect to different
 213 amounts of training data. To prepare the dataset, we integrate the MM model for 1250k
 214 days in total. The computational cost is low, requiring 1 CPU core and approximately
 215 11 hours. We select the first n days (with n ranging from 1k to 1000k) to create the train-
 216 ing data set, and always take the last 250k days for the test dataset. Thus all models
 217 can be fairly compared. The trajectory length and the corresponding number of nascent
 218 blocked state states are shown in Tab. 1. The likelihood q of forming a blocking event
 219 varies depending on different persistence thresholds D . This dependence relationship is
 220 illustrated in Tab. 2.

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.

Threshold	$Y = 1$	$Y = 0$	Positive rate
≥ 5 days	18748	69642	0.212
≥ 7 days	8522	79868	0.096
≥ 9 days	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 days, thus evolving into a blocking event under these respective thresholds, while $Y = 0$ denotes the number that don’t make it to the threshold.

221

5.2 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:

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

222

223

224

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 predicts high probability for positive events, and low probability for negative events.

225

226

227

228

229

230

231

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 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, which would both be zero.

232

233

234

235

236

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.

237

5.3 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

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}, \quad (10)$$

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}, \quad (11)$$

238
239
240
241
242

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

243
244
245
246
247
248
249

More informally, if the method forecasts that an event will occur, the precision measures the fraction of times this forecast is correct. The recall, on the other hand, is the fraction of the successfully forecasted events of all the positive events. If, regardless of the system state, one randomly predicts events with the climatological mean rate, in which an overall fraction p of the data labels are True, 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 better than climatological rate.

250
251
252
253
254
255
256

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

$$\text{Overall performance} = \text{Precision} + \text{Recall}. \quad (12)$$

257
258
259
260
261
262
263

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 naive criteria 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.

264

5.4 Performance and early stopping technique

265
266
267
268
269
270
271
272
273

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

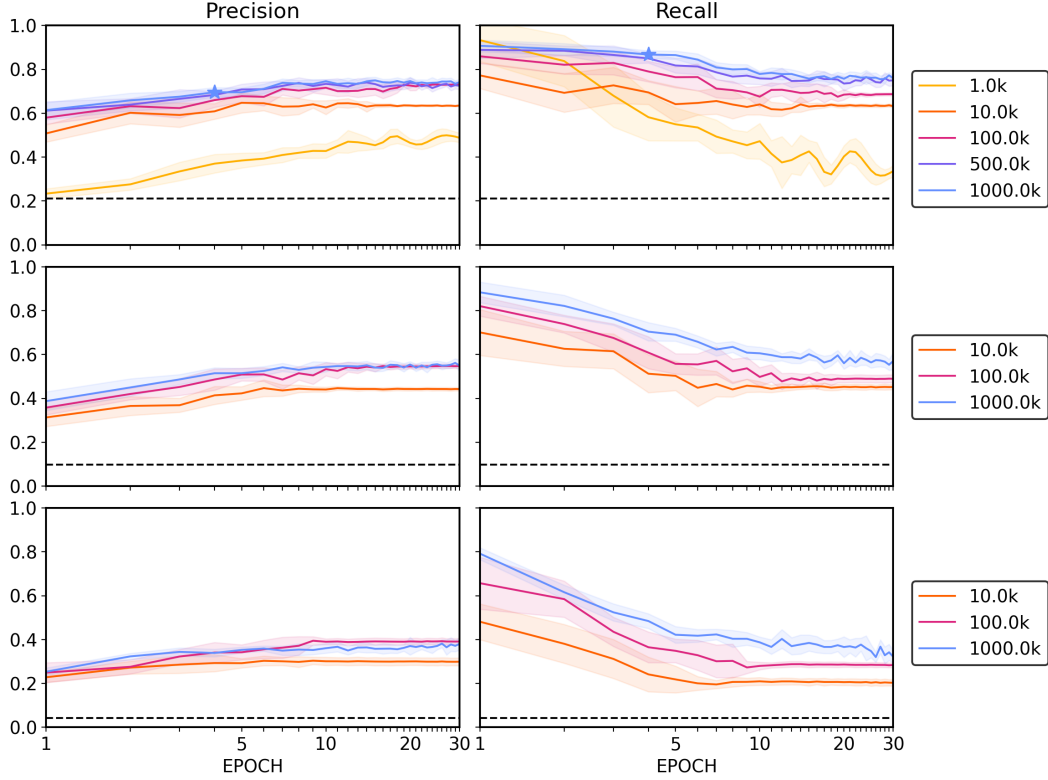


Figure 4. Top: Precision and recall results for models trained on data sets of varying sizes with $D = 5$. The dashed black line is the predicted recall and precision from the climatology computed using the largest data set. The blue stars indicate precision=0.70 and recall=0.87. Middle: Same results for $D = 7$. Bottom: Same results for $D = 9$. Fewer curves are displayed for $D = 7$ and $D = 9$ for the sake of clarity.

274 recall are approximately equal, but this is not necessarily the ideal stopping time (Miloshevich
 275 et al., 2023).

276 Applying the definition of best performance in Eq. (12), the “best” CNN is obtained
 277 by training on the full data set of 1000k days for 4 epochs, indicated by the star in Fig. 4.
 278 It achieves precision of 0.70 and recall of 0.87, exhibiting significant predictive power over
 279 the climatological mean prediction (the black dashed line with value 0.21). Therefore,
 280 we use it for further analysis in Sec. 6.

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

290 Fig. 4 also shows the results for higher persistence thresholds, $D = 7$ and 9. These
 291 thresholds correspond to rarer events, and even with the longest trajectory of 1000k days,
 292 the precision and recall curves suffer for two reasons. First, as seen from Tab. 2, the num-

ber 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 days block, and less than 1 in 20 into a 9 day 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 converges to about 0.35. While this is only half the values achieved by the CNN in the $D = 5$ case, this is almost 10 times the climatological values of precision and recall in that case. 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, thus appears more sensitive to class imbalance. More data gives the network more true positive cases to learn from, appearing to help overcome this challenge.

The low precision and recall values for smaller data sets (1k and 10k) does not bode well for training our CNN on ERA5 data, which will be discussed in detail in Section 8. 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, achieving a high recall is the most difficult with limited data, and it is with this metric that transfer learning will have the largest impact.

6 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.

6.1 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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$ (such as the conditional probability function q in Eq. 9), SHAP assigns an importance value ϕ_i to each feature x_i of the argument $\mathbf{x} \in \mathbb{R}^d$, which combine additively:

$$f(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] + \sum_{i=1}^d \phi_i(f, \mathbf{x}). \quad (13)$$

With no knowledge of \mathbf{x} , the optimal prediction of f (in a mean-square sense) is the climatological average over the distribution of \mathbf{x} : $\mathbb{E}[f(\mathbf{x})]$. SHAP values quantify how much is gained beyond this baseline by incorporating information from each component i of \mathbf{x} . The SHAP values $\phi_i(f, \mathbf{x})$ are unique for each sample of \mathbf{x} , but features i for which $|\phi_i(f, \mathbf{x})|$ are large for most \mathbf{x} (that is, a large SHAP value on average) can be singled out as important, or useful, for the prediction of $f(\mathbf{x})$. SHAP values possess advanta-

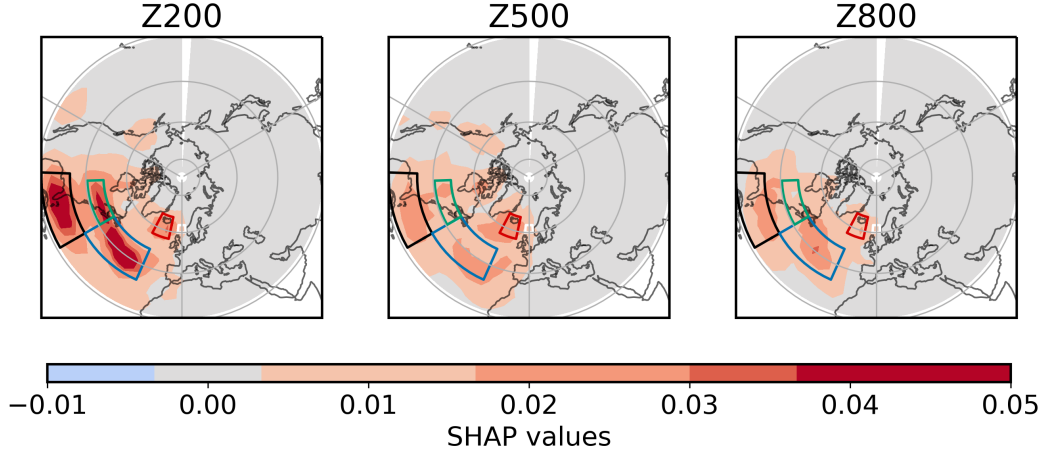


Figure 5. Composite maps of SHAP values, $\bar{\phi}$, of geopotential height at 200, 500, and 800 hPa, for true positive cases, i.e., when the CNN accurately forecasts a persistent blocking event. The unit is the probability of a positive forecast per feature (see equation 13), indicating the average incremental increase in the CNN’s confidence that the nascent blocked state will evolve into a persistent blocking event, given knowledge of Z at a given location and pressure. 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) .

336 geous theoretical properties as well, and we refer the reader to Lundberg and Lee (2017)
 337 for a detailed theoretical analysis. In this study, SHAP values are computed using the
 338 Python package Deep SHAP. The function $f(\mathbf{x})$ is taken as the estimated conditional
 339 probability $\hat{q}(\mathbf{x})$ computed by the CNN, i.e., the probability, according to the CNN, that
 340 the blocked state will extend $\geq D$ days, leading to a blocking event.

341 6.2 Results

342 Fig. 5 shows the composite of SHAP values for true positive data. Because few nascent
 343 blocks persist for $D = 5, 7, \text{ or } 9$, the climatological probability of a persistent event $\mathbb{E}[\hat{q}(\mathbf{x})] =$
 344 $0.21, 0.096, \text{ and } 0.044$, respectively. For our CNN to call a positive event, we require the
 345 conditional forecast probability $\hat{q}(\mathbf{x})$ to be larger than 0.5. Hence a positive (negative)
 346 value of $\phi_i(\hat{q}, \mathbf{x})$ indicates that knowing the geopotential height anomaly at this level and
 347 location increases (decreases) the likelihood of a positive event. Therefore, the shading
 348 in Fig. 5 can be interpreted as the average influence of each grid point for the CNN to
 349 successfully predict a long-lasting blocking event.

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

356 The first thing to notice is that anomalies upstream from the blocking region (to
 357 the west) are more valuable for predicting the persistence of the blocked state. More-
 358 over, the commonality among different pressure levels reflects the relatively barotropic

359 nature of the MM model. In general, however, the CNN prediction relies most on the
 360 upper level flow (200 hPa).

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

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

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

384 7 Building a sparse model: Logistic regression

385 To substantiate the importance of the regions highlighted by the CNN in predic-
 386 tion, we constructed a sparse model. The success of this model reveals that a small set
 387 of well-chosen variables and a model with a simple structure can recover a sizeable por-
 388 tion of the predictability. We computed the local mean of Z_{200} , Z_{500} , Z_{800} for each
 389 of the four rectangles shown in Fig. 5, resulting in 12 time series. We then applied lo-
 390 gistic regression with different combinations of these 12 features. The results for the sparse
 391 models with the best predictive skill within models of 1 to 5 dimensions on the test set
 392 are illustrated in Fig. 7(a). The horizontal axis denotes the variable combinations that
 393 achieve the predictive scores shown in the figure.

394 We draw three key conclusions from Fig. 7(a). First, to predict the persistence of
 395 a blocked state, the best one-dimensional feature is Z_{200} in region 1, upstream over Florida
 396 and the Gulf, not Z_{500} in region 4, the Z -field nearest to the blocking region we focus
 397 on. Second, the combination of Z_{200} in region 1, Z_{500} in region 4 forms a two-dimension
 398 model (shown in Fig. 7(b)) that already recovers a recall value of 0.75 – it captures three
 399 quarters of all blocking events – with a precision of 0.44, twice the climatological rate.
 400 The precision and recall of the full CNN, however, are 0.87 and 0.70. This leads us to
 401 the third key message: the large discrepancy in precision between CNN and logistic re-
 402 gression. Even with 5 predictors, the precision of our sparse model is only 0.5.

403 The poor precision indicates that the sparse model makes too many false positive
 404 predictions. This could suggest that the decay of the Atlantic blocked state is a more
 405 nonlinear dynamical phenomenon, which cannot be modeled as a simple linear statisti-
 406 cal model. A CNN can capture these nonlinearities more effectively than sparse regres-
 407 sion, which is consistent with previous research which found North Atlantic blocks are
 408 associated with nonlinear processes (Evans & Black, 2003). It could also indicate that

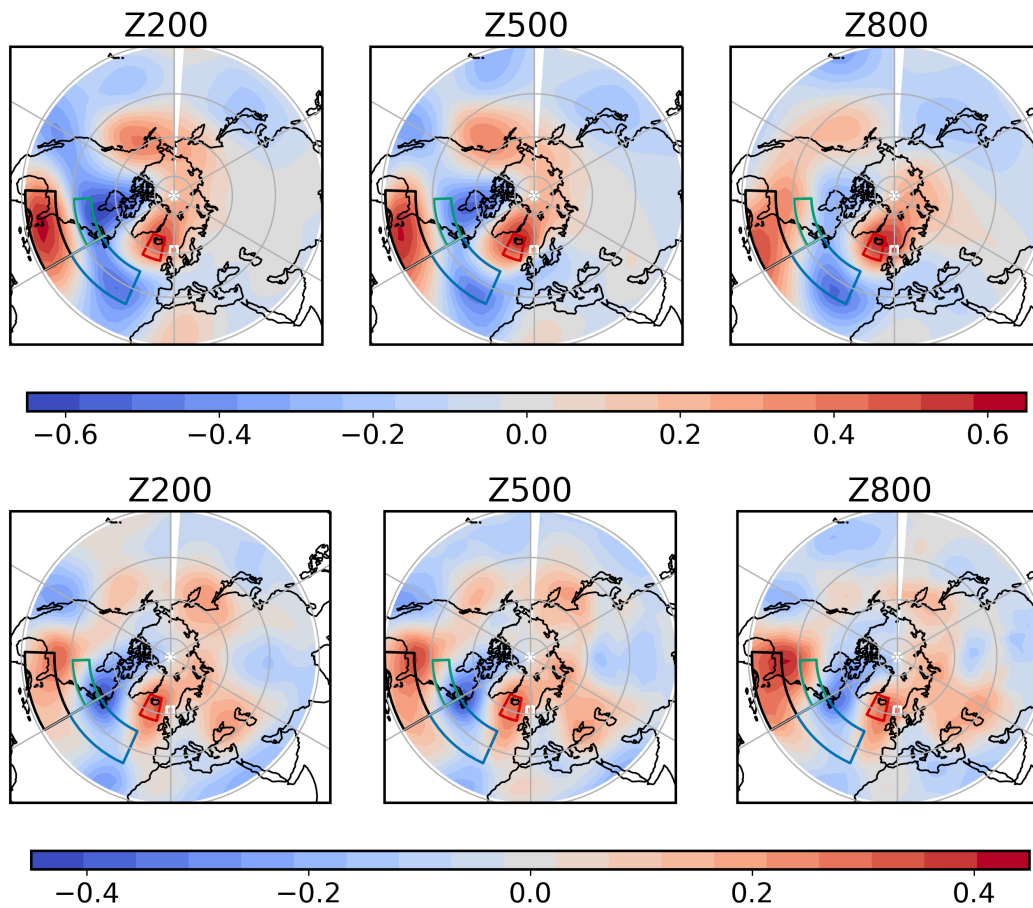


Figure 6. Average states of nascent blocking states that evolve into persistent blocking events ($T = 1, y = 1$) of (top row) MM dataset and (bottom row) ERA5. The colorbar represents values of geopotential height anomalies normalized by the standard deviation at each location and height.

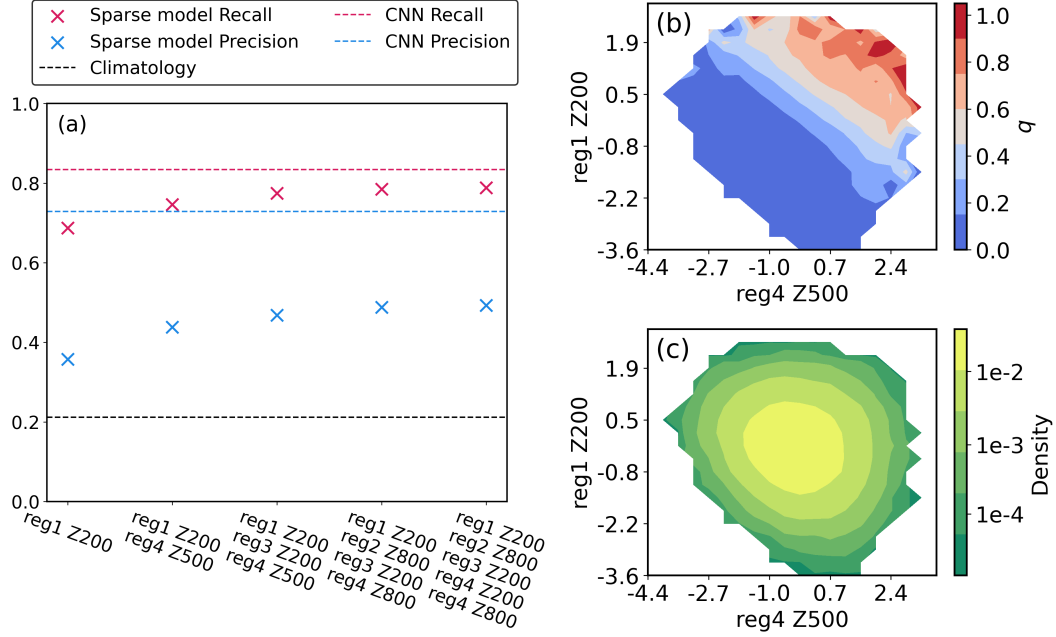


Figure 7. (a): Sparse model predictive skill on the test data set. The horizontal axis represents the dimension d of the sparse model from 1 to 5, with labels showing the combination of variables (“R1” = “region 1”) that achieves the best predictive skill among all combinations of d variables. The (+) and (−) indicate the sign of the coefficient before the variable in the logistic regression. (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.

409 more subtle features outside these 4 centers (and variation within these regions) are im-
 410 portant. Fig. 5 indicates that the CNN uses information across all of the North Atlantic,
 411 eastern North America, and even off the west coast of the US, to make skillful predic-
 412 tions.

413 To explore the effectiveness of the two-dimensional sparse model, we visualized the
 414 conditional probability of a block persisting, q , projected onto this simple subspace (shown
 415 in Fig. 7(b)). Introduced in Eq. (9), q quantifies the probability that the system will evolve
 416 into a persistent blocking event before the flow becomes unblocked. For example, the light-
 417 est pink region, corresponding to $q \approx 0.5$ indicates that if, at the onset of blocking, Z
 418 at 200 hPa over region 1 (Florida) is particularly high or Z at 500 hPa in region 4 (Ice-
 419 land) is abnormally high, the system has a roughly 50% chance of evolving into a per-
 420 sistent block, more than double the climatological rate of 21%. In the red region at the
 421 top right, where both of these regions exhibit abnormally high pressure, the odds of a
 422 persistent block increase to near 100%.

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

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

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

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

438 8 Extending to ERA5 Using Transfer Learning

439 Given sufficient data, it was possible to construct a CNN that skillfully forecasts
440 the maintenance of blocking events in the MM model. ERA5 December, January and
441 February (DJF) data from 1940-2022 exhibit only 273 nascent blocked states in our At-
442 lantic region of focus. A significant degradation in performance was evident in Fig. 4 when
443 we restricted the amount of training data from the MM model, the drop unfortunately
444 occurring in the data regime available in reanalysis. The curve associated with the tra-
445 jectory of 10k days (699 nascent blocked states) plateaued at lower values for both the
446 precision and recall. With only 1k days (63 nascent blocked states) performance was poor,
447 and the learning unstable, oscillating significantly across epochs.

448 The class imbalance between $Y = 0$ and $Y = 1$ adds to the difficulty (see Tab. 3),
449 particularly when longer blocks are considered. An extreme example is the set of block-
450 ing events that last ≥ 9 days: there are only 18 such events ($Y = 1$) in the reanalysis
451 record out of 273 data points. Such a small sample of positive data can hardly support
452 any meaningful training, and makes it impractical to get meaningful uncertainty bounds
453 on performance. In a standard training-test data split with a ratio of 90:10, only around
454 2 positive events typically fall in the test set, making it challenging to robustly assess
455 the skill.

456 When training on the limited number of events in the reanalysis, a CNN can more
457 easily suffer from overfitting, where the network uses ‘noise’ (unrelated features) to clas-
458 sify blocking events. Overfitting can be diagnosed when the performance on the test set
459 diverges from the training set. Yang and Gerber (submitted) found that the oversam-
460 pling strategy used so far in this study was more prone to overfitting than a weighted
461 loss function strategy (Johnson & Khoshgoftaar, 2019). With this latter strategy, one
462 emphasizes the rare class (in our case, positive events) by increasing its weight in the loss
463 function. In our remaining experiments, we weighted positive and negative events inverse
464 to their occurrence rate.

465 8.1 Direct training

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

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 day 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 size of samples. The average performance more reliably quantifies the potential skill. 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 day events, where only 13% of nascent blocks evolve into 7+ day events. Again, the average CNN modestly beat 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 Bernouli 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.

8.2 Transfer learning

Transfer learning (TL) has found broad application in atmospheric science, such as detecting gravity waves (González et al., 2022), improving extreme heatwave forecast in climate model (Jacques-Dumas et al., 2022), subgrid-scale (SGS) models (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

518 & Thuerey, 2021) or a dynamical model (Mu et al., 2020) with the observational record
 519 to improve medium-range weather forecasting and ENSO prediction.

520 In this section, we applied TL to leverage our MM dataset to predict events in the
 521 reanalysis data. As a quasi-geostrophic model, MM has complexity between full climate (Rasp
 522 & Thuerey, 2021) and low order (Mu et al., 2020) models used in previous transfer learn-
 523 ing studies. The overall process is to first ‘pre-train’ a CNN the MM model dataset, learn-
 524 ing to capture the characteristic features of blocking. While significantly simplified, the
 525 MM model is skillful in representing atmospheric variability (Lucarini & Gritsun, 2020),
 526 but more importantly provides extensive positive and negative cases to learn from, sup-
 527 porting optimal CNN training, as demonstrated in Sec. 5. After pre-training, our CNN
 528 is then fine-tuned on the ERA5 dataset, where the weights are modified to account for
 529 biases in the MM model, and the parameter scales are calibrated.

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

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

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

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

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

568 We still evaluate the overall performance by Eq. (12). Focusing first on $D = 5$ events,
 569 the best mean performance with DT is a precision of 0.45 and recall of 0.61, which is re-
 570 alized at Epoch 3. With TL, we achieve an average performance with a similar preci-
 571 sion of 0.45 and higher recall 0.82 (at Epoch 4). A noticeable advantage of TL is the sig-
 572 nificantly reduced variance in recall compared to DT, indicating TL’s superior robust-
 573 ness in prediction, attributed to its enhanced capacity for capturing predictive features.
 574 For $D = 7$ day events, the best mean performance with DT is a precision of 0.21 and
 575 recall of 0.48, achieved after 3 epochs. TL, however, achieves a precision of 0.22 and re-
 576 call of 0.76 at Epoch 6.

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

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

592 8.3 What has transfer learning learned?

593 When we show ERA5 events to CNNs first trained on the MM dataset, what ex-
 594 actly is the CNN learning to improve the recall? For example, do the key geographical
 595 regions and levels (Fig. 5) retain the same level of significance? It is reasonable to ex-
 596 pect that this might not be the case. In the MM dataset, the duration of the Atlantic
 597 blockings could be related to upstream flow, specifically to the structure of the wave train
 598 at the blocking onset. The mechanism for blocking in the real world is more complicated,
 599 and the correlated pattern may shift, intensify, and/or weaken. To address these ques-
 600 tions, we compare the SHAP values of the pre-trained CNNs when directly applied to
 601 ERA5 (i.e., without fine-tuning step) to the SHAP values of the CNN after 4 epochs of
 602 fine-tuning, as shown in row *a* and row *b* of Fig. 10. The most evident difference after
 603 fine-tuning is a decrease in the amplitude of the SHAP values. This is because the cli-
 604 matological rate of positive blocking events in ERA5 is higher: almost 1/3 of nascent blocked
 605 states persist for 5 days in ERA5, compared to about 1/5 in MM. As the expected frac-
 606 tion of events is larger, $\hat{q}(\mathbf{x}) - \mathbb{E}[\hat{q}(\mathbf{x})]$ from equation (13) will be smaller, and the SHAP
 607 value increments $\phi_i(\hat{q}, \mathbf{x})$ will tend to be smaller. It is the sum of the SHAP values that
 608 build up the probability for a $Y = 1$ prediction; for a more likely event, one does not
 609 need to build up the probability as much, so fine-tuning quickly adjusts the weights.

610 To assess the more subtle change in the relative contribution of each feature on the
 611 predicted result after transfer learning, we show the difference in the normalized com-
 612 posite map $\Delta\phi$ in row *d* of Fig. 10. $\Delta\phi$ is defined for each input i (i.e., geopotential height
 613 Z at a particular latitude, longitude, and pressure level) by $\Delta\phi_i \equiv \max\left(\frac{\bar{\phi}_i^{\text{TL}}}{\frac{1}{d} \sum_{j=1}^d \bar{\phi}_j^{\text{TL}}}, 0\right) -$
 614 $\max\left(\frac{\bar{\phi}_i}{\frac{1}{d} \sum_{j=1}^d \bar{\phi}_j}, 0\right)$. The maximum function is used to avoid spurious negative SHAP
 615 values, which should not arise in a composite of true positive events, as discussed in the
 616 context of Fig. 5. The normalization makes the total integral of the SHAP values the

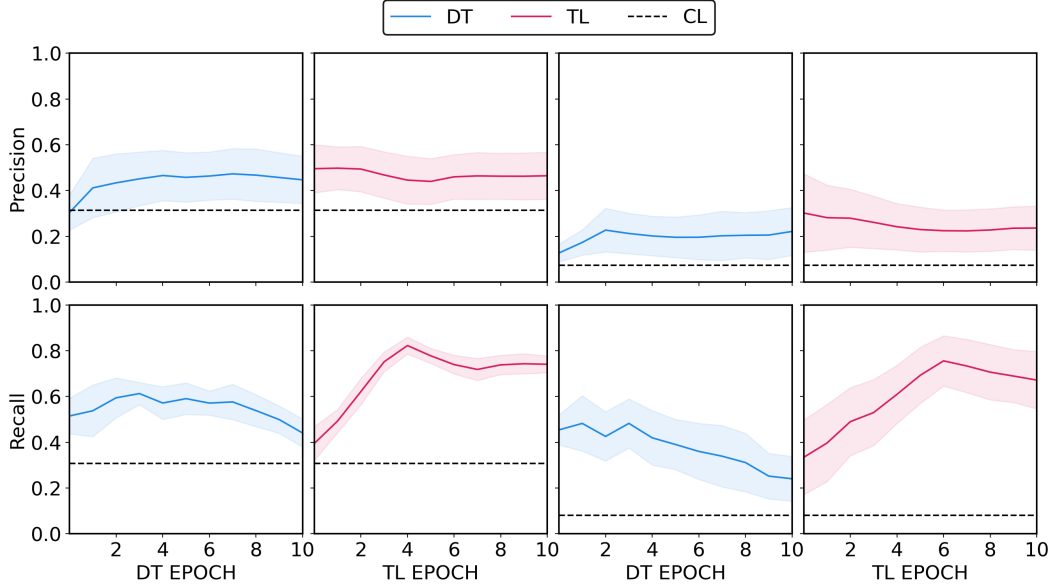


Figure 8. Comparison of CNN forecast skill between direct training (DT, blue) and transfer learning (TL, red). The top row shows the precision and the bottom row, the recall, as a function of training epoch of DT (columns 1 and 3) and fine-tuning epoch of TL (columns 2 and 4). The black dashed line indicates the climatological event rate p . The left two columns show the results for $D = 5$ (standard blocking events) and the right two columns show the results for $D = 7$ (longer blocking events). The shading shows a two-standard deviation uncertainty bound, as detailed in the text.

617 same for both cases, so that one can focus on where the CNN is using information, as
 618 opposed to the overall reduction of the SHAP values driven by the difference in rates.

619 The “normalized” SHAP values increase mainly in region 4 (the region right around
 620 the block), and additionally over Quebec and Atlantic Canada, a region less used for pre-
 621 dictions with the MM model. The SHAP values decrease in a relative sense over regions
 622 1 (Florida and the Gulf), 2 (North Atlantic Ocean), 3 (northeastern North America),
 623 and central North America. This change in relative importance reveals a general de-emphasis
 624 of the regions farther upstream and an increased emphasis on regions more immediately
 625 upstream. This indicates that while it is still upstream information that is most impor-
 626 tant for predicting a persistent blocking state in ERA5, the structure and westward ex-
 627 tension of the wave train has changed.

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

636 The SHAP values change over Quebec, capturing this shift in the wave train, but
 637 overall the CNN seems to shift to more local information with transfer learning. We spec-
 638 ulated that the dry, quasi-geostrophic MM model overemphasizes long range teleconnec-
 639 tions. It only captures deformation scale dynamics, and this only at low resolution, and

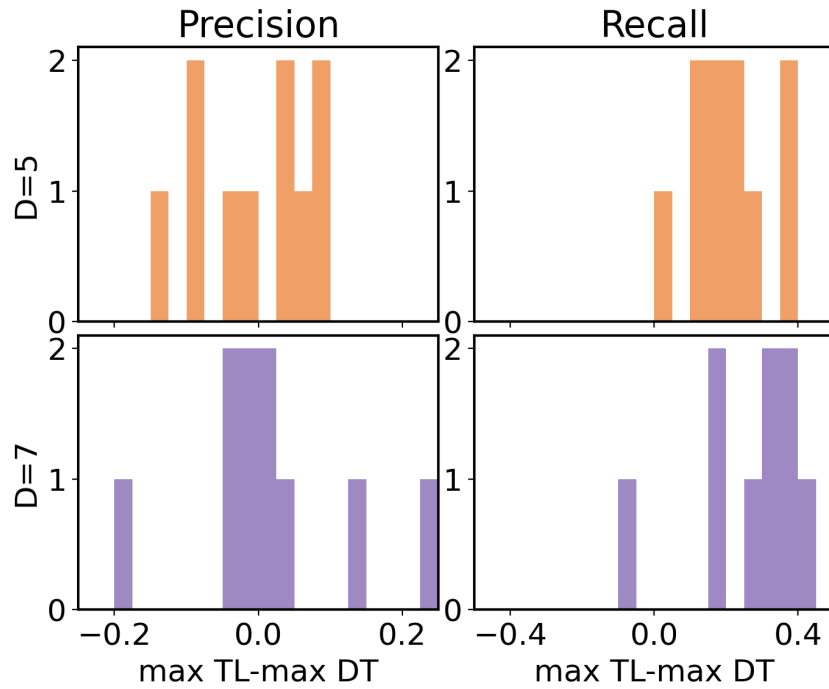


Figure 9. Histograms of the performance gap between the best performing CNNs obtained with transfer learning versus the best performing CNNs obtained with direct training, for (left) precision and (right) recall. The top panels are for 5 day events and the lower panels are for 7 day events. “Best performing” 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.

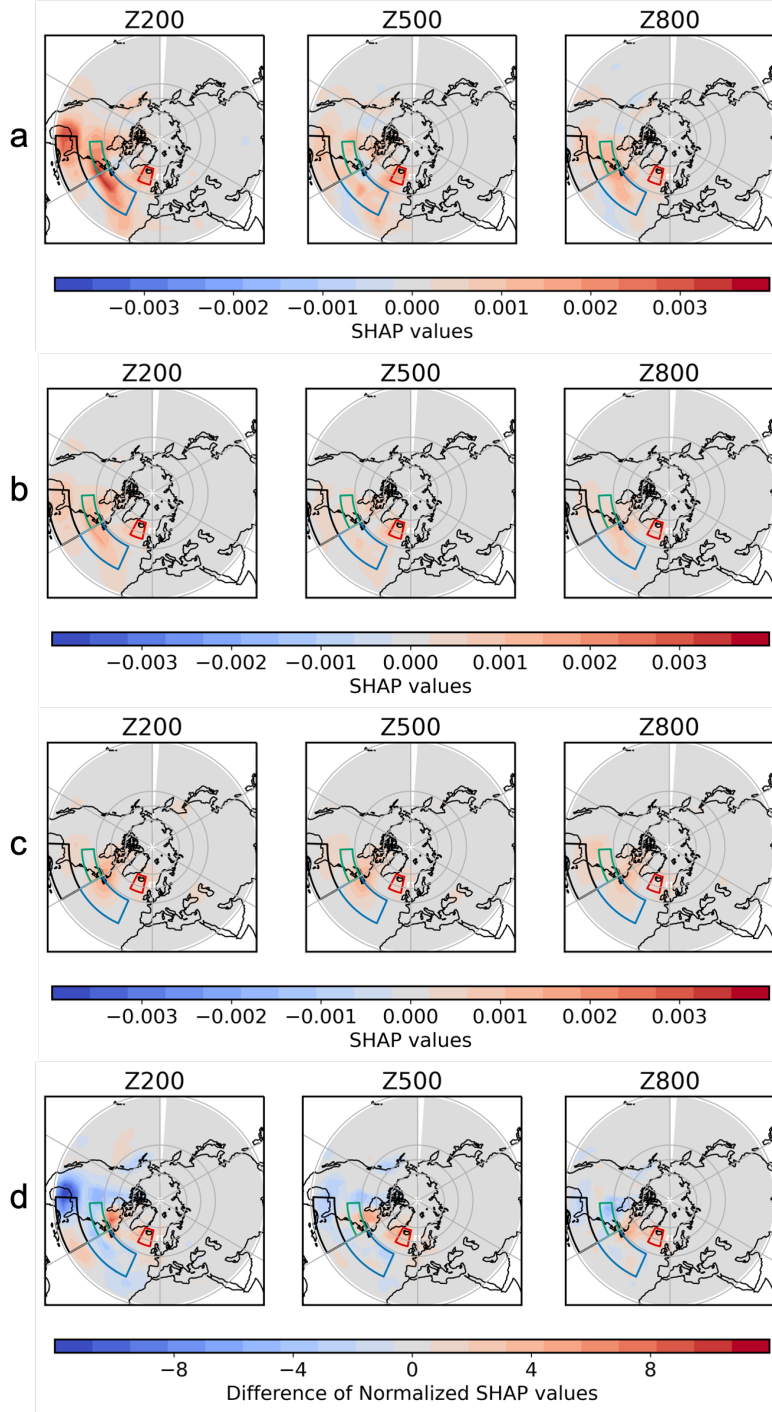


Figure 10. Rows 1 through 4 are composite maps of SHAP values, $\bar{\phi}$, for geopotential height (200, 500, and 800 hPa), averaged over true positive predictions of blocking events in ERA5 by the CNNs listed below. This is the same quantity shown in Fig.5, but now applied to ERA5 events. Row *a* shows $\bar{\phi}^{\text{MM}}$ for the pre-trained CNNs before transfer learning (i.e., networks that have only learned from MM, but applied to ERA5). Row *b*: $\bar{\phi}^{\text{TL}}$ of these pre-trained CNNs after fine-tuning. Row *c*: $\bar{\phi}^{\text{DT}}$ of CNNs directly trained on ERA5 dataset (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.

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.

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.

9 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, XAI methods 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 XAI (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

690 the training data was limited, particularly for the recall statistic. Through transfer learn-
 691 ing, we leverage the abundance of data generated by simplified dynamical models to en-
 692 hance real-world forecasting. By pre-training a CNN on the MM model dataset and re-
 693 training the deepest layer on the ERA5 dataset, the recall was improved by 34% com-
 694 pared to a CNN developed with direct training alone for 5 day events, and over 50% for
 695 more extreme 7 day events, without any loss of precision.

696 In addition to advancing predictive skill, transfer learning in combination with SHAP
 697 analysis allowed us to compare the predictive features between weather systems in ERA5
 698 and the idealized quasigeostrophic model. The bottom row of Fig. 6 reveals biases in the
 699 MM model, which appears overly dependent on upstream features over Florida and the
 700 Gulf of Mexico relative to blocks in ERA5. This approach provides a new angle of how
 701 a machine learning approach could guide the diagnosis and quantification of model bi-
 702 ases. This said, the success of transfer learning results underscore the MM model’s abil-
 703 ity, despite its simplicity, to capture features that are important for predicting the per-
 704 sistence of blocked states in the real world. We believe that greater strides could be made
 705 by pre-training on a more advanced climate model, or even hindcasts in the subseasonal-
 706 to-seasonal (S2S) data set (Vitart et al., 2017; Finkel et al., 2023).

707 The methods presented here are not limited to the context of blocking events, and
 708 can be generalized to the study of other challenging natural phenomena, especially in
 709 scenarios where data may be limited, and the potential influencing factors are complex.
 710 An immediate future goal is to push further on the physical and dynamical mechanisms
 711 that causes the differences in prediction mechanisms for ERA5 and MM model. Another
 712 goal is to adapt the present approach to investigate the statistical behavior and mech-
 713 anisms for the onset of the blocking events.

714 Open Research Section

715 The code for computing SHAP values, transfer learning and producing plots is pub-
 716 licly available in the Github repository at https://github.com/hzhang-math/Blocking_SHAP_TL.
 717

718 Acknowledgments

719 We thank Valerio Lucarini and Andrey Gritsun for sharing their Marshall-Molteni For-
 720 tran code. This work was supported by the Army Research Office, grant number W911NF-
 721 22-2-0124. EPG acknowledges support from the National Science Foundation through
 722 award OAC-2004572. J. F. is supported through the MIT Climate Grand Challenge on
 723 Weather and Climate Extremes, and the Virtual Earth Systems Research Institute (VESRI)
 724 at Schmidt Sciences.

725 References

- 726 Barnes, E. A., & Hartmann, D. L. (2010). Dynamical feedbacks and the persis-
 727 tence of the nao. *Journal of the Atmospheric Sciences*, *67*(3), 851 - 865.
 728 Retrieved from [https://journals.ametsoc.org/view/journals/atsc/67/3/](https://journals.ametsoc.org/view/journals/atsc/67/3/2009jas3193.1.xml)
 729 [2009jas3193.1.xml](https://journals.ametsoc.org/view/journals/atsc/67/3/2009jas3193.1.xml) doi: 10.1175/2009JAS3193.1
- 730 Berckmans, J., Woollings, T., Demory, M.-E., Vidale, P.-L., & Roberts, M. (2013).
 731 Atmospheric blocking in a high resolution climate model: influences of mean
 732 state, orography and eddy forcing. *Atmospheric Science Letters*, *14*(1), 34–40.
- 733 Chan, P.-W., Hassanzadeh, P., & Kuang, Z. (2019). Evaluating Indices of Blocking
 734 Anticyclones in Terms of Their Linear Relations With Surface Hot Extremes.
 735 *Geophysical Research Letters*, *46*(9), 4904–4912. Retrieved from [https://](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019GL083307)
 736 agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019GL083307 doi:
 737 <https://doi.org/10.1029/2019GL083307>

- 738 Charney, J. G., & DeVore, J. G. (1979). Multiple flow equilibria in the atmosphere
739 and blocking. *Journal of Atmospheric Sciences*, *36*(7), 1205–1216.
- 740 Conover, W. J. (1999). *Practical nonparametric statistics* (Vol. 350). John Wiley &
741 sons.
- 742 Davini, P., & D’Andrea, F. (2020). From CMIP3 to CMIP6: Northern Hemisphere
743 Atmospheric Blocking Simulation in Present and Future Climate. *Journal of*
744 *Climate*, *33*(23), 10021 - 10038. Retrieved from [https://journals.ametsoc](https://journals.ametsoc.org/view/journals/clim/33/23/jcliD190862.xml)
745 [.org/view/journals/clim/33/23/jcliD190862.xml](https://journals.ametsoc.org/view/journals/clim/33/23/jcliD190862.xml) doi: [https://doi.org/10](https://doi.org/10.1175/JCLI-D-19-0862.1)
746 [.1175/JCLI-D-19-0862.1](https://doi.org/10.1175/JCLI-D-19-0862.1)
- 747 Davini, P., & D’Andrea, F. (2016). Northern Hemisphere atmospheric blocking rep-
748 resentation in global climate models: twenty years of improvements? *Journal*
749 *of Climate*, *29*(24), 8823–8840.
- 750 Dikshit, A., & Pradhan, B. (2021). Explainable AI in drought forecasting. *Ma-*
751 *chine Learning with Applications*, *6*, 100192. Retrieved from [https://](https://www.sciencedirect.com/science/article/pii/S2666827021000967)
752 www.sciencedirect.com/science/article/pii/S2666827021000967 doi:
753 <https://doi.org/10.1016/j.mlwa.2021.100192>
- 754 Dole, R. M., & Gordon, N. D. (1983). Persistent anomalies of the extratropical
755 Northern Hemisphere wintertime circulation: Geographical distribution and
756 regional persistence characteristics. *Monthly Weather Review*, *111*(8), 1567–
757 1586.
- 758 d’Andrea, F., Tibaldi, S., Blackburn, M., Boer, G., Déqué, M., Dix, M., ... others
759 (1998). Northern Hemisphere atmospheric blocking as simulated by 15 atmo-
760 spheric general circulation models in the period 1979–1988. *Climate Dynamics*,
761 *14*, 385–407.
- 762 Evans, K. J., & Black, R. X. (2003). Piecewise tendency diagnosis of weather
763 regime transitions. *Journal of the Atmospheric Sciences*, *60*(16), 1941 -
764 1959. Retrieved from [https://journals.ametsoc.org/view/journals/](https://journals.ametsoc.org/view/journals/atasc/60/16/1520-0469_2003_060_1941_ptdowr_2.0.co_2.xml)
765 [atasc/60/16/1520-0469_2003_060_1941_ptdowr_2.0.co_2.xml](https://journals.ametsoc.org/view/journals/atasc/60/16/1520-0469_2003_060_1941_ptdowr_2.0.co_2.xml) doi:
766 [10.1175/1520-0469\(2003\)060<1941:PTDOWR>2.0.CO;2](https://doi.org/10.1175/1520-0469(2003)060<1941:PTDOWR>2.0.CO;2)
- 767 Ferranti, L., Corti, S., & Janousek, M. (2015). Flow-dependent verification of
768 the ECMWF ensemble over the Euro-Atlantic sector. *Quarterly Journal*
769 *of the Royal Meteorological Society*, *141*(688), 916–924. Retrieved from
770 <https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2411>
771 doi: <https://doi.org/10.1002/qj.2411>
- 772 Finkel, J., Webber, R. J., Gerber, E. P., Abbot, D. S., & Weare, J. (2023).
773 Data-Driven Transition Path Analysis Yields a Statistical Understanding
774 of Sudden Stratospheric Warming Events in an Idealized Model. *Journal*
775 *of the Atmospheric Sciences*, *80*(2), 519 - 534. Retrieved from [https://](https://journals.ametsoc.org/view/journals/atasc/80/2/JAS-D-21-0213.1.xml)
776 journals.ametsoc.org/view/journals/atasc/80/2/JAS-D-21-0213.1.xml
777 doi: <https://doi.org/10.1175/JAS-D-21-0213.1>
- 778 González, J. L., Chapman, T., Chen, K., Nguyen, H., Chambers, L., Mostafa, S. A.,
779 ... Yue, J. (2022). Atmospheric Gravity Wave Detection Using Transfer
780 Learning Techniques. In *2022 IEEE/ACM International Conference on Big*
781 *Data Computing, Applications and Technologies (BDCAT)* (p. 128–137). doi:
782 [10.1109/BDCAT56447.2022.00023](https://doi.org/10.1109/BDCAT56447.2022.00023)
- 783 Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
784 (<http://www.deeplearningbook.org>)
- 785 Guo, Y., Wu, X., Qing, C., Su, C., Yang, Q., & Wang, Z. (2022). Blind Restora-
786 tion of Images Distorted by Atmospheric Turbulence Based on Deep Trans-
787 fer Learning. *Photonics*, *9*(8). Retrieved from [https://www.mdpi.com/](https://www.mdpi.com/2304-6732/9/8/582)
788 [2304-6732/9/8/582](https://www.mdpi.com/2304-6732/9/8/582) doi: [10.3390/photonics9080582](https://doi.org/10.3390/photonics9080582)
- 789 Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019, Sep 01). Deep learning for multi-year
790 ENSO forecasts. *Nature*, *573*(7775), 568–572. Retrieved from [https://doi](https://doi.org/10.1038/s41586-019-1559-7)
791 [.org/10.1038/s41586-019-1559-7](https://doi.org/10.1038/s41586-019-1559-7) doi: [10.1038/s41586-019-1559-7](https://doi.org/10.1038/s41586-019-1559-7)
- 792 Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater,

- 793 J., ... Thépaut, J.-N. (2020). The era5 global reanalysis. *Quarterly Jour-*
794 *nal of the Royal Meteorological Society*, 146(730), 1999-2049. Retrieved from
795 <https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803> doi:
796 <https://doi.org/10.1002/qj.3803>
- 797 Hoskins, B. J., James, I. N., & White, G. H. (1983, July). The Shape, Propagation
798 and Mean-Flow Interaction of Large-Scale Weather Systems. *Journal of At-*
799 *mospheric Sciences*, 40(7), 1595-1612. doi: 10.1175/1520-0469(1983)040<1595:
800 TSPAMF>2.0.CO;2
- 801 Hussain, M., Bird, J. J., & Faria, D. R. (2019). A study on cnn transfer learning for
802 image classification. In *Advances in Computational Intelligence Systems: Con-*
803 *tributions Presented at the 18th UK Workshop on Computational Intelligence,*
804 *September 5-7, 2018, Nottingham, UK* (pp. 191-202).
- 805 Jacques-Dumas, V., Ragone, F., Borgnat, P., Abry, P., & Bouchet, F. (2022). Deep
806 learning-based extreme heatwave forecast. *Frontiers in Climate*, 4.
- 807 Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class
808 imbalance. *Journal of Big Data*, 6(1), 1-54.
- 809 Kautz, L.-A., Martius, O., Pfahl, S., Pinto, J. G., Ramos, A. M., Sousa, P. M., &
810 Woollings, T. (2022). Atmospheric blocking and weather extremes over the
811 Euro-Atlantic sector – a review. *Weather and Climate Dynamics*, 3(1), 305-
812 336. Retrieved from <https://wcd.copernicus.org/articles/3/305/2022/>
813 doi: 10.5194/wcd-3-305-2022
- 814 Labe, Z. M., & Barnes, E. A. (2021). Detecting Climate Signals Using Ex-
815 plainable AI With Single-Forcing Large Ensembles. *Journal of Ad-*
816 *vances in Modeling Earth Systems*, 13(6), e2021MS002464. Retrieved
817 from [https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002464)
818 [2021MS002464](https://doi.org/10.1029/2021MS002464) (e2021MS002464 2021MS002464) doi: [https://doi.org/10.1029/](https://doi.org/10.1029/2021MS002464)
819 [2021MS002464](https://doi.org/10.1029/2021MS002464)
- 820 Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A re-
821 view of machine learning interpretability methods. *Entropy*, 23(1), 18.
- 822 Lipovetsky, S., & Conklin, M. (2001). Analysis of regression in game theory ap-
823 proach. *Applied Stochastic Models in Business and Industry*, 17(4), 319-330.
824 Retrieved from [https://onlinelibrary.wiley.com/doi/abs/10.1002/](https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.446)
825 [asmb.446](https://doi.org/10.1002/asmb.446) doi: <https://doi.org/10.1002/asmb.446>
- 826 Liu, Y., Racah, E., Prabhat, Correa, J., Khosrowshahi, A., Lavers, D., ... Collins,
827 W. (2016). *Application of Deep Convolutional Neural Networks for Detecting*
828 *Extreme Weather in Climate Datasets*.
- 829 Lucarini, V., & Gritsun, A. (2020). A new mathematical framework for atmospheric
830 blocking events. *Climate Dynamics*, 54(1-2), 575-598.
- 831 Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model pre-
832 dictions. *Advances in neural information processing systems*, 30.
- 833 Lupo, A. R. (2021). Atmospheric blocking events: a review. *Annals of the New*
834 *York Academy of Sciences*, 1504(1), 5-24. Retrieved from [https://nyaspubs](https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/nyas.14557)
835 [.onlinelibrary.wiley.com/doi/abs/10.1111/nyas.14557](https://doi.org/10.1111/nyas.14557) doi: [https://doi](https://doi.org/10.1111/nyas.14557)
836 [.org/10.1111/nyas.14557](https://doi.org/10.1111/nyas.14557)
- 837 Lupo, A. R., Mokhov, I. I., Akperov, M. G., Chernokulsky, A. V., Athar, H., et al.
838 (2012). A dynamic analysis of the role of the planetary-and synoptic-scale
839 in the summer of 2010 blocking episodes over the European part of Russia.
840 *Advances in Meteorology*, 2012.
- 841 Malmgren-Hansen, D., Nielsen, A. A., Laparra, V., & Valls, G. C. (2018). Transfer
842 Learning with Convolutional Networks for Atmospheric Parameter Retrieval.
843 In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing*
844 *Symposium* (p. 2111-2114). doi: 10.1109/IGARSS.2018.8518097
- 845 Marshall, J., & Molteni, F. (1993). Toward a dynamical understanding of planetary-
846 scale flow regimes. *Journal of the atmospheric sciences*, 50(12), 1792-1818.
- 847 Matsueda, M. (2009). Blocking Predictability in Operational Medium-Range Ensem-

- 848 ble Forecasts. *SOLA*, 5, 113-116. doi: 10.2151/sola.2009-029
- 849 McWilliams, J. C. (1980). An application of equivalent modons to atmospheric
850 blocking. *Dynamics of Atmospheres and Oceans*, 5(1), 43-66. Retrieved from
851 <https://www.sciencedirect.com/science/article/pii/037702658090010X>
852 doi: [https://doi.org/10.1016/0377-0265\(80\)90010-X](https://doi.org/10.1016/0377-0265(80)90010-X)
- 853 Michelangeli, P.-A., & Vautard, R. (1998). The dynamics of Euro-Atlantic blocking
854 onsets. *Quarterly Journal of the Royal Meteorological Society*, 124(548), 1045–
855 1070.
- 856 Miloshevich, G., Cozian, B., Abry, P., Borgnat, P., & Bouchet, F. (2023, Apr).
857 Probabilistic forecasts of extreme heatwaves using convolutional neural net-
858 works in a regime of lack of data. *Phys. Rev. Fluids*, 8, 040501. Retrieved
859 from <https://link.aps.org/doi/10.1103/PhysRevFluids.8.040501> doi:
860 10.1103/PhysRevFluids.8.040501
- 861 Mu, B., Ma, S., Yuan, S., & Xu, H. (2020). Applying convolutional lstm net-
862 work to predict el niño events: Transfer learning from the data of dynamical
863 model and observation. In *2020 IEEE 10th International Conference on Elec-
864 tronics Information and Emergency Communication (ICEIEC)* (p. 215-219). doi:
865 10.1109/ICEIEC49280.2020.9152317
- 866 Mullen, S. L. (1987). Transient eddy forcing of blocking flows. *Journal of the Atmo-
867 spheric Sciences*, 44(1), 3–22.
- 868 Pelly, J. L., & Hoskins, B. J. (2003). A new perspective on blocking. *Journal of the
869 atmospheric sciences*, 60(5), 743–755.
- 870 Rampal, N., Gibson, P. B., Sood, A., Stuart, S., Fauchereau, N. C., Brandolino,
871 C., ... Meyers, T. (2022). High-resolution downscaling with interpretable
872 deep learning: Rainfall extremes over New Zealand. *Weather and Climate
873 Extremes*, 38, 100525. Retrieved from [https://www.sciencedirect.com/
874 science/article/pii/S2212094722001049](https://www.sciencedirect.com/science/article/pii/S2212094722001049) doi: [https://doi.org/10.1016/
875 j.wace.2022.100525](https://doi.org/10.1016/j.wace.2022.100525)
- 876 Rasp, S., & Thuerey, N. (2021). Data-Driven Medium-Range Weather Predic-
877 tion With a Resnet Pretrained on Climate Simulations: A New Model for
878 WeatherBench. *Journal of Advances in Modeling Earth Systems*, 13(2),
879 e2020MS002405. Retrieved from [https://agupubs.onlinelibrary.wiley
880 .com/doi/abs/10.1029/2020MS002405](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002405) (e2020MS002405 2020MS002405) doi:
881 <https://doi.org/10.1029/2020MS002405>
- 882 Rex, D. F. (1950). Blocking Action in the Middle Troposphere and its Effect
883 upon Regional Climate. *Tellus*, 2(3), 196-211. Retrieved from [https://
884 onlinelibrary.wiley.com/doi/abs/10.1111/j.2153-3490.1950.tb00331.x](https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2153-3490.1950.tb00331.x)
885 doi: <https://doi.org/10.1111/j.2153-3490.1950.tb00331.x>
- 886 Rudy, S. H., & Sapsis, T. P. (2023). Output-weighted and relative entropy loss
887 functions for deep learning precursors of extreme events. *Physica D: Nonlinear
888 Phenomena*, 443, 133570.
- 889 Scaife, A. A., Woollings, T., Knight, J., Martin, G., & Hinton, T. (2010). Atmo-
890 spheric blocking and mean biases in climate models. *Journal of Climate*,
891 23(23), 6143–6152.
- 892 Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features
893 through propagating activation differences. In *International conference on ma-
894 chine learning* (pp. 3145–3153).
- 895 Shutts, G. (1983). The propagation of eddies in diffluent jetstreams: Eddy vortic-
896 ity forcing of ‘blocking’ flow fields. *Quarterly Journal of the Royal Meteorologi-
897 cal Society*, 109(462), 737–761.
- 898 Silva, S. J., Keller, C. A., & Hardin, J. (2022). Using an explainable machine
899 learning approach to characterize Earth System model errors: Application of
900 SHAP analysis to modeling lightning flash occurrence. *Journal of Advances in
901 Modeling Earth Systems*, 14(4), e2021MS002881.
- 902 Subel, A., Chattopadhyay, A., Guan, Y., & Hassanzadeh, P. (2021). Data-driven

- 903 subgrid-scale modeling of forced Burgers turbulence using deep learning with
 904 generalization to higher Reynolds numbers via transfer learning. *Physics of*
 905 *Fluids*, 33(3).
- 906 Talo, M., Baloglu, U. B., Yıldırım, Ö., & Acharya, U. R. (2019). Application of
 907 deep transfer learning for automated brain abnormality classification using MR
 908 images. *Cognitive Systems Research*, 54, 176–188.
- 909 Tibaldi, S., & Molteni, F. (1990). On the operational predictability of blocking. *Tel-*
 910 *lus A*, 42(3), 343–365.
- 911 Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., ...
 912 Zhang, L. (2017). The Subseasonal to Seasonal (S2S) Prediction Project
 913 Database. *Bulletin of the American Meteorological Society*, 98(1), 163 - 173.
 914 Retrieved from [https://journals.ametsoc.org/view/journals/bams/98/1/](https://journals.ametsoc.org/view/journals/bams/98/1/bams-d-16-0017.1.xml)
 915 [bams-d-16-0017.1.xml](https://journals.ametsoc.org/view/journals/bams/98/1/bams-d-16-0017.1.xml) doi: 10.1175/BAMS-D-16-0017.1
- 916 Woollings, T., Barriopedro, D., Methven, J., Son, S.-W., Martius, O., Harvey, B., ...
 917 Seneviratne, S. (2018, Sep 01). Blocking and its Response to Climate Change.
 918 *Current Climate Change Reports*.
- 919 Yang, M., Luo, D., Li, C., Yao, Y., Li, X., & Chen, X. (2021). Influ-
 920 ence of Atmospheric Blocking on Storm Track Activity Over the
 921 North Pacific During Boreal Winter. *Geophysical Research Let-*
 922 *ters*, 48(17), e2021GL093863. Retrieved 2023-08-03, from [https://](https://onlinelibrary.wiley.com/doi/abs/10.1029/2021GL093863)
 923 onlinelibrary.wiley.com/doi/abs/10.1029/2021GL093863 (_eprint:
 924 <https://onlinelibrary.wiley.com/doi/pdf/10.1029/2021GL093863>) doi:
 925 10.1029/2021GL093863
- 926 Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are fea-
 927 tures in deep neural networks? *Advances in neural information processing sys-*
 928 *tems*, 27.
- 929 Zappa, G., Masato, G., Shaffrey, L., Woollings, T., & Hodges, K. (2014a). Linking
 930 Northern Hemisphere blocking and storm track biases in the CMIP5 climate
 931 models. *Geophysical Research Letters*, 41(1), 135–139.
- 932 Zappa, G., Masato, G., Shaffrey, L., Woollings, T., & Hodges, K. (2014b). Linking
 933 Northern Hemisphere blocking and storm track biases in the CMIP5 climate
 934 models. *Geophysical Research Letters*, 41(1), 135–139. Retrieved 2023-08-02,
 935 from <https://onlinelibrary.wiley.com/doi/abs/10.1002/2013GL058480>
 936 (_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/2013GL058480>) doi:
 937 10.1002/2013GL058480