Environmental Data Science (2022), 1: e17, 1–9
doi:10.1017/eds.2022.19

CAMBRIDGE
UNIVERSITY PRESS

APPLICATION PAPER

Modeling stratospheric polar vortex variation and identifying vortex extremes using explainable machine learning

Zheng Wu[1,*], Tom Beucler[2], Enikő Székely[3], William T. Ball[4] and Daniela I.V. Domeisen[1,2]

[1]Institute for Atmospheric and Climate Science, ETH Zürich, Zürich, Switzerland
[2]Institute of Earth Surface Dynamics, University of Lausanne, Lausanne, Switzerland
[3]Swiss Data Science Center, ETH Zürich and EPFL, Lausanne, Switzerland
[4]Department of Geoscience and Remote Sensing, TU Delft, Delft, The Netherlands
*Corresponding author. E-mail: zheng.wu@env.ethz.ch

Received: 29 September 2022; **Accepted:** 25 October 2022

Keywords: Explainable neural networks; feature selection; polar vortex extreme events; prediction; stratospheric variation

Abstract

The winter stratospheric polar vortex (SPV) exhibits considerable variability in magnitude and structure, which can result in extreme SPV events. These extremes can subsequently influence weather in the troposphere from weeks to months and thus are important sources of surface predictability. However, the predictability of the SPV extreme events is limited to 1–2 weeks in state-of-the-art prediction systems. Longer predictability timescales of SPV would strongly benefit long-range surface prediction. One potential option for extending predictability timescales is the use of machine learning (ML). However, it is often unclear which predictors and patterns are important for ML models to make a successful prediction. Here we use explainable multiple linear regressions (MLRs) and an explainable artificial neural network (ANN) framework to model SPV variations and identify one type of extreme SPV events called sudden stratospheric warmings. We employ a NN attribution method to propagate the ANN's decision-making process backward and uncover feature importance in the predictors. The feature importance of the input is consistent with the known precursors for extreme SPV events. This consistency provides confidence that ANNs can extract reliable and physically meaningful indicators for the prediction of the SPV. In addition, our study shows a simple MLR model can predict the SPV daily variations using sequential feature selection, which provides hints for the connections between the input features and the SPV variations. Our results indicate the potential of explainable ML techniques in predicting stratospheric variability and extreme events, and in searching for potential precursors for these events on extended-range timescales.

Impact Statement

This study explores the application of explainable machine learning methods and their attribution methods in modeling stratospheric variations and identifying stratospheric extreme events, which can be used to improve the forecast skill of surface weather. A simple linear regression model is built to predict the stratospheric variations using the feature selection method and the neural network is able to identify extreme stratospheric events. More importantly, the neural network attribution technique provides insights that the reasoning behind the decision-making process of neural networks is interpretable and reliable. This study sheds light on the potential of explainable neural networks in searching for opportunities for skillful prediction of stratospheric extreme events and, by extension, surface weather beyond weekly time scales.

https://doi.org/10.1017/eds.2022.19 Published online by Cambridge University Press

1. Motivation

The stratospheric polar vortex (SPV) is a strong circumpolar westerly wind band in the polar stratosphere that forms in fall, decays in spring, and exhibits strong variability in both magnitude and zonal wave structure during winter. The SPV variability in mid-winter mainly depends on the interaction between planetary waves and the background mean flow in the stratosphere (McIntyre, 1982). Anomalously strong (weak) planetary wave driving in the lower stratosphere can result in extremely weak (strong) SPV strength. Major sudden stratospheric warmings (SSWs) are one type of extremely weak SPV events, during which the SPV breaks down. Extreme SPV events like SSWs can subsequently influence the tropospheric circulation and weather from weeks to months, leading to extreme weather over North America and Europe (Domeisen and Butler, 2020), such as cold air outbreaks and extreme snowfall (King et al., 2019). Therefore, SSW events are thought to be an important source of predictability on weekly to monthly timescales over the Northern Hemisphere (NH) mid- and high-latitudes (Karpechko et al., 2017). Improving the predictability of SSW events may thus help to enhance the forecast skill in the troposphere (Domeisen et al., 2020b). However, the predictability of SSW events is limited to around 1–2 weeks in state-of-the-art prediction systems (Domeisen et al., 2020a).

Previous studies used different statistical models to predict stratospheric variability and identify SSW events at timescales longer than 1 week, such as multiple linear regression (MLR) and fully connected neural networks (NNs), and found that a well-trained NN can exhibit promising skill in the prediction of SPV variations (e.g., Blume and Matthes, 2012; Peng et al., 2021). However, given the multiple variables with different lead times used in these studies and the nonlinear structure of NNs, the key factors and features that the NNs use to make the prediction are not clear. Before exploring the application of NNs in extended-range prediction of SPV variations and extreme events, we need to gain confidence that NNs use physically meaningful features in the input variables to produce their predictions. On the other hand, it is generally easier to interpret the predictions made by a linear regression model. However, it is difficult to understand the essential factors for the prediction given too many input variables and the intercorrelation across these variables. Therefore, using as little feature as possible to make regression models more transparent can aid the dynamical understanding of the predictors and the SPV variations.

Recent studies have highlighted that extended-range weather prediction opportunities can be identified by employing artificial neural networks (ANNs) together with their visualization techniques (Barnes et al., 2020; Mayer and Barnes, 2021). We, therefore, use an explainable neural network framework to predict the SSW events. In this study, we use geopotential height and background zonal mean flow in the lower stratosphere as predictors based on dynamical knowledge of the atmosphere. Our results show that the features in the input that are relevant for NNs to make the prediction are consistent with the precursor patterns found by dynamical analyses.

Details about data, the statistical models, and attribution methods are given in Section 2. Section 3 presents the outcomes of SPV prediction using different models and feature visualization methods. Finally, the results are summarized and discussed in Section 4.

2. Data and Methods

2.1. Data

The data used in this study are from the European Centre for Medium-Range Weather Forecasts (ECMWF) Interim reanalysis (ERA-interim, Dee et al., 2011). We use daily mean zonal mean zonal wind at 60°N and 10 hPa ($U10_{60}$) to represent the SPV strength. The zonal wind at 50 hPa ($U50$) north of 30°N is used to represent the extratropical stratospheric background state. The geopotential height deviations from the zonal-mean at 100 hPa ($Z100$) north of 30°N are used to represent the wave driving in the lower stratosphere. Given that the strong SPV variability and extremes are concentrated around winter, we focus on this particular season (from November to March) over the period 1979–2018. Daily climatologies are removed from all the variables used in the study. A low-pass Lanczos filtering (Duchon, 1979) is applied to the $U10_{60}$ daily anomalies to filter out timescales of less than 10 days as we are not

https://doi.org/10.1017/eds.2022.19 Published online by Cambridge University Press

interested in high-frequency $U10_{60}$ fluctuations. We then apply principal component (PC) analysis on $Z100$ and $U50$, respectively. As we discuss in the following sections, we only use a limited number of modes to represent or reconstruct $Z100$ and $U50$ in the statistical models.

2.2. Statistical models

In this study, we aim to (1) predict $U10_{60}$ daily anomalies, which is a regression problem and (2) predict SSW events, which is a classification problem. Given the different goals of the two problems, we describe the different statistical models used in the two tasks below.

2.2.1. Regression model

We choose $Z100$ and $U50$ as predictors to model the $U10_{60}$ daily anomalies. The first 200 PC time series of $Z100$ (explaining 99% of the variance) averaged from 10 to 1 days before the day for which $U10_{60}$ is predicted (target day) and the first 100 PC time series of $U50$ (explaining 99% of the variance) averaged from 20 to 11 days before the target day are used as input features (total 300 modes) to predict the low-pass filtered standardized $U10_{60}$ daily anomalies. These two physical fields and the different time periods are chosen as the background mean flow serves as a waveguide for the upward wave propagation and the wave driving in the lower stratosphere affects the SPV strength (e.g., Bancalá et al., 2012). The input data format is a 2D tensor of shape (samples, modes) and the output is the $U10_{60}$ anomaly, which we use to reconstruct the full 1D time series.

We use a MLR model from sklearn (Pedregosa et al., 2011) for this regression task. The data are separated into training (32 out of 40 winters) and test data (8 out of 40 winters). The MLR is trained 60 times with different combinations of training and test data, which are randomly separated and shuffled. The test data are only used to evaluate the performance of the trained MLR model. Three performance metrics are used for the regression model over the training and test data: the mean absolute error (MAE), the correlation coefficient (r), and the coefficient of determination (R^2). For this simple regression task, we found that ANNs did not outperform MLRs (not shown), and therefore focus on our MLR model in the results sections.

2.2.2. Classification model

In this task, we aim to classify a winter into a normal winter or a winter with an SSW event. We build an ANN model using $Z100$ as input. We use the first 20 modes of $Z100$ (explaining 76% of the variance) to reconstruct $Z100$. In our case, using too many modes does not help to improve the accuracy of the classification because it leads to strong overfitting. The reconstructed $Z100$ field is then averaged from 10 to 1 days before the onset day of SSWs and any 10 successive normal winter days as input for the ANN. The $Z100$ with this time period is chosen since we will compare the feature importance highlighted by the ANN's visualization tool with previous dynamical-based studies that used the same variables at the same time lags (Martius et al., 2009). The input data format is a 3D tensor of shape (samples, longitudes, and latitudes). The output of the ANN are probabilities that indicate whether the winter exhibits an SSW event or not. Given that we have a relatively small sample size, we only separate the data into training (35 samples) and test (8 samples) data.

We implement the ANNs using Keras (Chollet, 2018). The ANN architecture consists of two hidden layers with 10 neurons each and using the rectified linear activation function (ReLU). Its final layer contains two neurons and a softmax activation. The ANN uses categorical cross entropy for loss function and is trained for 50 epochs with the SGD optimizer with a fixed learning rate of 0.001. We manually tune the number of neurons of the hidden layers and other hyperparameters such as the activation function and the number of epochs. The ANN with the current choices of hyperparameters produces high accuracy as shown in Table 2. Given that the goal of the study is to understand the relevance of the input features for the ANN to make its decision, we do not perform a thorough search of hyperparameters to obtain a perfect ANN model. We use the accuracy and F1 score as the performance metrics. The ANN is trained 60 times

https://doi.org/10.1017/eds.2022.19 Published online by Cambridge University Press

with random initialization and on different subset of data. Meanwhile, for each of the 60 experiments, we apply a logistic regression on the same data as a baseline.

2.3. Feature selection and visualization

The goal of this study is to extract the input features that are important for the ML models to predict the SPV variations and SSW events. To this end, we use two different methods to inspect the feature importance for the regression and classification tasks, respectively.

2.3.1. Sequential feature selection

In the regression task, even though R^2 increases with more modes used in the MLR model, we aim to make the model as transparent as possible. To this end, we use a forward sequential feature selection (SFS) from the sklearn.feature selection module (Ferri et al., 1994) to select 5 modes out of the total 300 modes. Starting with no selected feature, the forward SFS searches one feature at a time to maximize a cross-validated score, where R^2 is used in this study. Once that first feature is selected, the procedure is repeated by searching for the next best feature and adding to the set of selected features until the desired number of selected features (five modes/features) is reached.

2.3.2. Layerwise relevance propagation

In the classification task, since the ANN is trained to learn from the input features to make accurate predictions, interpreting the relationship learned by the ANN can provide insight into the stratospheric circulation and its coupling with the troposphere. Here we employ a visualization technique called Layerwise Relevance Propagation (LRP) (Bach et al., 2015) to extract a heat map, which indicates the relevance of each input feature to the final prediction. After the ANN is trained to classify normal winters and winters with an SSW event, all data are passed through the final ANN model. Then LRP is implemented to take the highest probability between the two categories and to backward propagate the relevance from the output neuron to the input layers (Toms et al., 2020). Given that the input layer here is the reconstructed $Z100$ with its first 20 modes, the output of the LRP algorithm is a heat map with the same dimension as the input (lon × lat), identifying the key regions that are given higher relevance for each ANN's classification. We can obtain both the individual and the composite of heat maps for all correctly classified SSWs in all 60 experiments and compare the patterns in the heat maps with that identified dynamically in previous studies.

3. Results

3.1. Modeling SPV daily variations

The performance metrics of the MLR using 300 modes as predictors are shown in the left columns in Table 1. Comparing the values of the metrics between training and test data, the MLR using 300 modes has a strong overfitting problem. When we inspect the predicted standardized $U10_{60}$ daily anomalies (orange line) for the test data shown in Figure 1, the MLR attempts to predict high-frequency variability that does not always match the "true" variability from the ERA-interim (blue line). However, when we only use five modes identified from the SFS procedure, the predicted daily anomalies (green line) are more aligned with the "true" values and do not show strong and abrupt changes from day to day. The mean, 25th, and 75th percentile of MAE (R^2 and r) values of the MLR using selected modes (right columns in Table 1) are smaller (greater) on the test data than the values that using the MLR with all 300 modes, and there is no severe overfitting issue when comparing between the training and test data.

Given the different training data, the best five modes that the SFS procedure selects differ a bit among the 60 experiments. However, the first modes of both $Z100$ and $U50$ are selected for all experiments. The sixth and seventh modes of $U50$ are chosen 56 and 52 times out of the 60 experiments. The second and fifth modes of $Z100$ are selected 26 and 24 times, respectively, and it is noted that only one of these two

https://doi.org/10.1017/eds.2022.19 Published online by Cambridge University Press

Table 1. Performance metrics for the regression task of the MLR using 300 modes as predictors (left columns) and the MLR using the five best modes identified by the SFS procedure (right columns) over the training and test data.

	MLR using 300 modes		MLR using 5 modes	
Data set	Training	Test	Training	Test
Mean absolute error (MAE)	0.37 (0.37/0.38)	0.68 (0.66/0.71)	0.59 (0.58/0.6)	0.62 (0.6/0.64)
Coefficient of determination (R^2)	0.78 (0.77/0.78)	0.25 (0.16/0.34)	0.41 (0.39/0.43)	0.36 (0.32/0.44)
Correlation coefficient (r)	0.88 (0.88/0.89)	0.6 (0.56/0.65)	0.64 (0.63/0.65)	0.62 (0.58/0.67)

Note. The number shows the mean of the 60 experiments and the numbers in the parenthesis show the 25th percentile and 75th percentile of the 60 experiments, respectively.

Table 2. Performance metrics for the classification task of the ANN (left columns) and the logistic regression baseline (right columns) over the training and test data using 2D $Z100$ spatial patterns as input.

	Logistic regression (baseline)		ANN	
Data set	Training	Test	Training	Test
Accuracy	0.9 (0.86/0.97)	0.5 (0.38/0.63)	0.9 (0.86/0.97)	0.6 (0.5/0.7)
F1 score	1 (1/1)	0.5 (0.4/0.7)	0.9 (0.89/1.0)	0.6 (0.5/0.8)

Note. The number shows the mean of the 60 experiments and the numbers in the parenthesis show the 25th percentile and 75th percentile of the 60 experiments, respectively.

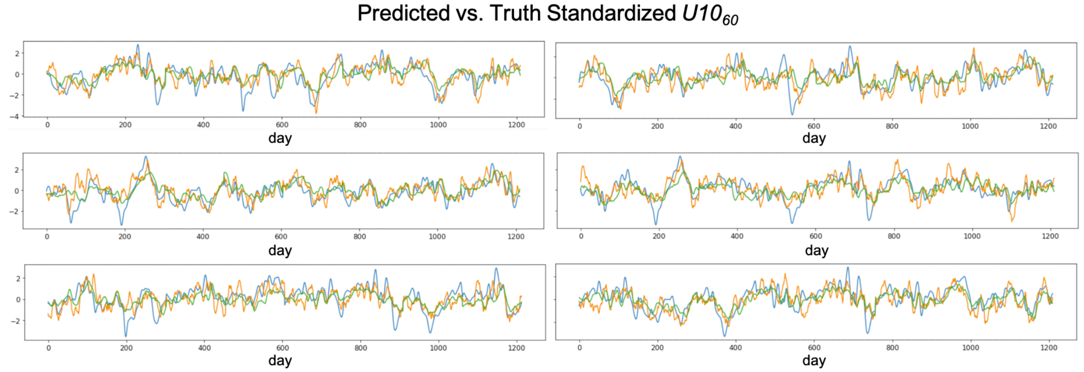
Predicted vs. Truth Standardized $U10_{60}$

Figure 1. *The low-pass filtered standardized $U10_{60}$ daily anomalies of the test data for the target ERA-interim values, which is our "truth" (blue), the predicted values obtained from the MLR using 300 modes (orange), and from the MLR using the five best modes selected by the SFS procedure (green). Each panel shows the results of one set of test data (8-year) from the 60 experiments. Note that the test data in the 60 experiments are different and here we only show 6 experiments as examples.*

modes is selected in the set of the best five features for most of the 60 experiments. The spatial patterns of these modes are shown in Figure 2. From the weights of the selected modes, the first modes of $Z100$ and $U50$ are more important than the other modes (not shown). The modes of $Z100$ identified by the SFS procedure are consistent with the dynamical understanding that waves with sufficiently small wavenumber (e.g., wave-1 and wave-2) can propagate upward into the stratosphere and interact with the mean wind flow (upper row of Figure 2). The spatial patterns of the $U50$ modes suggest not only the connection of

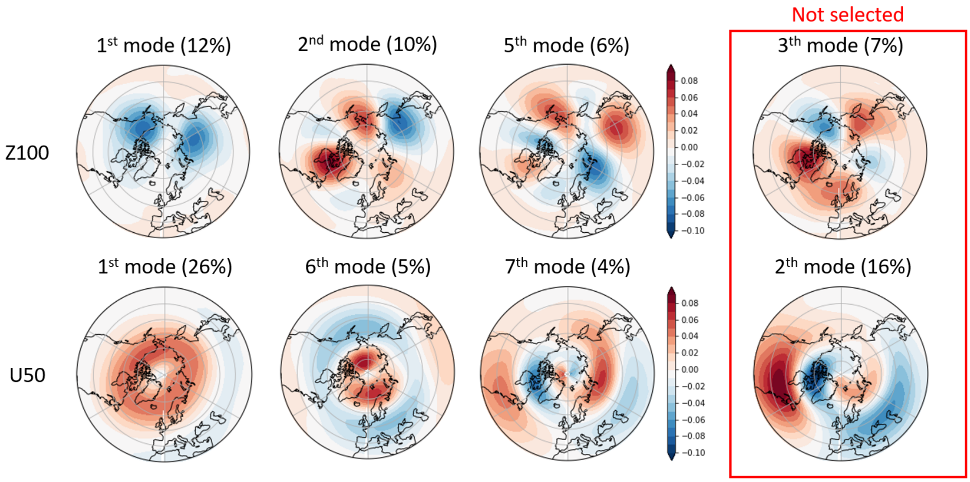

Figure 2. *The spatial patterns of the most selected modes of (upper row) Z*100 *and (bottom row) U*50 *by the forward SFS procedure. The numbers in the parenthesis show the explained variance by each mode. The two panels in the red box show the third mode of Z*100 *and second mode of U*50*, which represents larger variability of the variables but are rarely selected by SFS.*

the lower stratospheric wind to the SPV variations but also the influence on the wave propagation (bottom row of Figure 2). The sixth and seventh modes of $U50$ have similar weights with opposite signs, of which the sum shows positive anomalies from 30° to 50°N over eastern North America and from 50° to 70°N over Eurasia, and negative anomalies over the polar region. This structure is consistent with the literature showing that strong zonal-mean zonal winds over the extratropics in the lower stratosphere tends to lead the upward propagating planetary waves toward the equator and thus leading to a stronger SPV (Sigmond and Scinocca, 2010). It is interesting to note that the SFS does not only select modes which represent large variability of the variables as shown in the red box in Figure 2. Figure 2 suggests that the structures of the zonal asymmetries in the zonal wind are considered to be important for the SPV prediction by the SFS procedure. Since most of the previous studies using dynamical analysis focus on the zonal mean zonal wind, the spatial patterns identified here can potentially facilitate the understanding of the role of zonal asymmetries of the zonal wind in the interaction with the planetary waves and the SPV variations.

3.2. Identifying SSW events

In addition to predicting the daily variation of the SPV strength, we also predict the occurrence of the SPV extreme events, which can influence the tropospheric weather. We train an ANN model and a logistic regression baseline to classify the winters into normal winters and winters with SSWs using $Z100$, which directly influences the occurrence of SSWs (Bancalá et al., 2012). The performance metrics of both models are shown in Table 2. The mean of the performance metrics of the 60 experiments is better in the ANN models than in the logistic regression models and the interquartile range is smaller in the ANN models, indicating the performance of the ANN model is consistently better than that of the baseline.

To gain confidence in the prediction of the ANN model, we use LRP to extract and visualize the features of $Z100$ that the trained ANN models use to correctly identify the SSW events. The composite relevance heat maps for all correctly identified SSW events in all 60 experiments are shown in Figure 3a. The warmer colors indicate a larger relevance, corresponding to the key regions that the ANN uses to make its correct classification. The key regions of $Z100$ shown in Figure 3a are consistent with those found in previous studies (e.g., Figure 2 in Martius et al., 2009): increased wave amplitude over northeastern North America, northern Eurasia at 100 hPa. Figure 3b,c shows the heat maps of two

https://doi.org/10.1017/eds.2022.19 Published online by Cambridge University Press

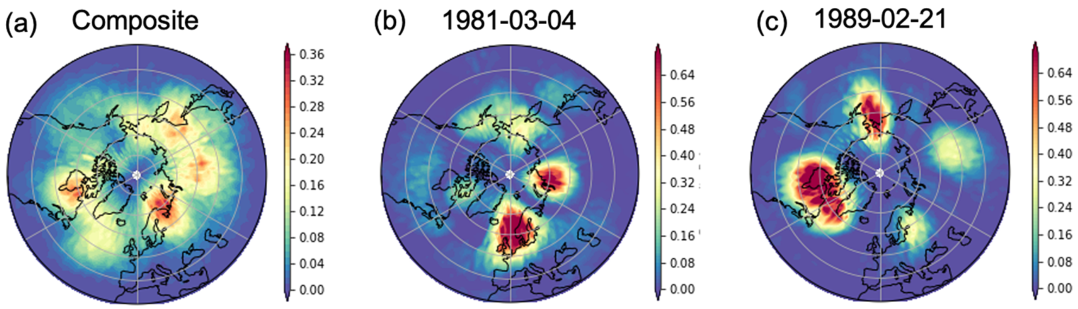

Figure 3. *The relevance heat maps of* Z100 *for correctly identified SSW events by the ANN models. (a) The composite of heat maps for correctly identified SSWs in all 60 experiments; (b) the heat map of the SSW event on March 4, 1981; and (c) the heat map of the SSW event on February 21, 1989.*

individual SSW events that occurred on March 4, 1981 and February 21, 1989, respectively. These two examples corresponded to blocking events before the SSW events, both in the Atlantic and in the Pacific (Martius et al., 2009). Large relevance is located in the corresponding regions in the heat maps of these two SSW events, indicating that the ANN is able to detect the anomalous strong wave activity as precursor for the SSW events. On the other hand, the individual heat maps also show high relevance in some other regions that do not correspond to the blocking (e.g., the Siberian region), which could potentially indicate so far undocumented precursors for SSWs. We also note that these XAI patterns may not be that robust without further selection criteria.

4. Conclusion

In this study, we use an MLR model together with the forward SFS method to predict the SPV variations and use the ANN model to predict the occurrence of SPV extreme events. The SFS approach allows us to simplify and improve the generalization ability of the MLR model by restricting the input vector to the most important features. Further analyzing these selected features fosters the dynamical understanding of the roles of zonal wave patterns and zonal asymmetries of the background wind in the SPV variations. Using the visualization method LRP, we demonstrate that the ANN model predict SSW events based on anomalously strong wave activity over various regions. In this study, we do not aim at a perfect model to predict the SPV variations and extreme events. Rather, we aim to train a statistical model that has sufficient predictive power to yield informative relevance. Therefore, we only search the hyperparameters of the ANN manually and keep the ANN very shallow (two layers) due to the limited number of samples. Both the ANN and the logistic regression model have overfitting issues, due to the small size of the training data. In a future study, the ANN model could be improved by obtaining more training data from climate models to build a deeper model. Extracting features that are learned from the data with minimal human intervention would help discover new precursors for prediction of these extreme events beyond weekly timescales. For example, the sea level pressure field could be fed to the NNs to predict the SPV extreme event at a lead time of several weeks and the attribution methods could be used to understand which features and regions in the input could potentially contribute to a successful long-term prediction. With the aid of the feature selection methods and explainable neural networks, our dynamical understanding of the conditions of the occurrence of extreme SPV events could be further improved by interpreting the relevance of the input learned from the data. Our study sheds light on the applicability of the explainable ML methods in identifying potential precursors for SPV variations and extreme events on extended-range timescales.

https://doi.org/10.1017/eds.2022.19 Published online by Cambridge University Press

Acknowledgments. Funding from the Swiss Data Science Center and the Swiss National Science Foundation is gratefully acknowledged. The authors would like to thank Guillaume Obozinski and Raphaël de Fondeville for their valuable suggestions and comments, and two anonymous reviewers for their helpful comments throughout the revision of the manuscript.

Author Contributions. Conceptualization: Z.W., D.I.V.D.; Data curation: Z.W., D.I.V.D.; Data visualization: Z.W.; Methodology: Z.W., T.B.; Writing—original draft: Z.W. All authors provided feedback for the manuscript and approved the final submitted draft.

Competing Interests. The authors declare no competing interests exist.

Data Availability Statement. ERA-Interim reanalysis was obtained from the ECMWF server (https://apps.ecmwf.int/datasets/data/interim-full-daily, last accessed May 2020).

Ethics Statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Funding Statement. This research was supported by grants from the Swiss Data Science Center within the project *EXPECT* (C18–08) to Z.W. and by the Swiss National Science Foundation through projects PP00P2_170523 and PP00P2_198896 to Z.W. and D.I.V.D.

Provenance. This article is part of the Climate Informatics 2022 proceedings and was accepted in *Environmental Data Science* on the basis of the Climate Informatics peer review process.

References

Bach S, **Binder A**, **Montavon G**, **Klauschen F**, **Müller KR and Samek W** (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One 10*(7), 1–46.

Bancalá S, **Krüger K and Giorgetta M** (2012) The preconditioning of major sudden stratospheric warmings. *Journal of Geophysical Research Atmospheres 117*(4), 1–12.

Barnes EA, **Mayer K**, **Toms B**, **Martin Z and Gordon E** (2020) Identifying opportunities for skillful weather prediction with interpretable neural networks, pp. 1–6.

Blume C and Matthes K (2012) Understanding and forecasting polar stratospheric variability with statistical models. *Atmospheric Chemistry and Physics 12*(13), 5691–5701.

Chollet F (2018) Keras: The python deep learning library. *Astrophysics Source Code Library*, ascl-1806.

Dee DP, **Uppala SM**, **Simmons AJ**, **Berrisford P**, **Poli P**, **Kobayashi S**, **Andrae U**, **Balmaseda MA**, **Balsamo G**, **Bauer P**, **Bechtold P**, **Beljaars AC**, **van de Berg L**, **Bidlot J**, **Bormann N**, **Delsol C**, **Dragani R**, **Fuentes M**, **Geer AJ**, **Haimberger L**, **Healy SB**, **Hersbach H**, **Hólm EV**, **Isaksen L**, **Kållberg P**, **Köhler M**, **Matricardi M**, **Mcnally AP**, **Monge-Sanz BM**, **Morcrette JJ**, **Park BK**, **Peubey C**, **de Rosnay P**, **Tavolato C**, **Thépaut JN and Vitart F** (2011) The ERA-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society 137*(656), 553–597.

Domeisen DIV and Butler AH (2020) Stratospheric drivers of extreme events at the Earth's surface. *Communications Earth & Environment 1*(59), 1–8.

Domeisen DIV, **Butler AH**, **Charlton-Perez AJ**, **Ayarzagüena B**, **Baldwin MP**, **Dunn-Sigouin E**, **Furtado JC**, **Garfinkel CI**, **Hitchcock P**, **Karpechko AY**, **Kim H**, **Knight J**, **Lang AL**, **Lim EP**, **Marshall A**, **Roff G**, **Schwartz C**, **Simpson IR**, **Son SW and Taguchi M** (2020) The role of the stratosphere in subseasonal to seasonal prediction: 2. Predictability of the stratosphere. *Journal of Geophysical Research: Atmospheres 125*(2), 1–20.

Domeisen DIV, **Butler AH**, **Charlton-Perez AJ**, **Ayarzagüena B**, **Baldwin MP**, **Dunn-Sigouin E**, **Furtado JC**, **Garfinkel CI**, **Hitchcock P**, **Karpechko AY**, **Kim H**, **Knight J**, **Lang AL**, **Lim EP**, **Marshall A**, **Roff G**, **Schwartz C**, **Simpson IR**, **Son SW and Taguchi M** (2020) The role of the stratosphere in subseasonal to seasonal prediction: 1. Predictability of the stratosphere. *Journal of Geophysical Research: Atmospheres 125*(2), 1–17.

Duchon CE (1979) Lanczos filtering in one and two dimensions. *Journal of Applied Meteorology and Climatology 18*(8), 1016–1022.

Ferri FJ, **Pudil P**, **Hatef M and Kittler J** (1994) Comparative study of techniques for large-scale feature selection. In *Machine Intelligence and Pattern Recognition*, vol. *16*. New York: Elsevier, pp. 403–413.

Karpechko AY, **Hitchcock P**, **Peters DHW and Schneidereit A** (2017) Predictability of downward propagation of major sudden stratospheric warmings. *Quarterly Journal of the Royal Meteorological Society 60*, 1459–1470.

King AD, **Butler AH**, **Jucker M**, **Earl NO and Rudeva I** (2019) Observed relationships between sudden stratospheric warmings and European climate extremes. *Journal of Geophysical Research: Atmospheres 124*(24), 13943–13961.

Martius O, **Polvani LM and Davies HC** (2009) Blocking precursors to stratospheric sudden warming events. *Geophysical Research Letters 36*(14), 1–5.

Mayer KJ and Barnes EA (2021) Subseasonal forecasts of opportunity identified by an explainable neural network. *Geophysical Research Letters 48*(10), 1–9.

https://doi.org/10.1017/eds.2022.19 Published online by Cambridge University Press

McIntyre E (1982) How well do we understand the dynamics of stratospheric warmings? *Journal of the Meteorological Society of Japan 60*(1), 37–65.

Pedregosa F, **Varoquaux G**, **Gramfort A**, **Michel V**, **Thirion B**, **Grisel O**, **Blondel M**, **Prettenhofer P**, **Weiss R**, **Dubourg V**, **Vanderplas J**, **Passos A**, **Cournapeau D**, **Brucher M**, **Perrot M and Duchesnay E** (2011) Scikit-learn: Machine learning in python. *Journal of Machine Learning Research 12*, 2825–2830.

Peng K, **Cao X**, **Liu B**, **Guo Y**, **Xiao C and Tian W** (2021) Polar vortex multi-day intensity prediction relying on new deep learning model: A combined convolution neural network with long short-term memory based on Gaussian smoothing method. *Entropy 23*(10), 1314.

Sigmond M and Scinocca JF (2010) The influence of the basic state on the northern hemisphere circulation response to climate change. *Journal of Climate 23*(6), 1434–1446.

Toms BA, **Barnes EA and Ebert-Uphoff I** (2020) Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems 12*(9), 1–20.

Cite this article: Wu Z. Beucler T. Székely E. Ball WT and Domeisen DIV, (2022). Modeling stratospheric polar vortex variation and identifying vortex extremes using explainable machine learning. *Environmental Data Science*, 1: e17. doi:10.1017/eds.2022.19