國立臺灣大學理學院大氣科學系
碩士論文
Department of Atmospheric Sciences
College of Science
National Taiwan University
Master Thesis

可解釋性深度學習方法於辨識平流層爆發性增溫事件之應用

Interpretable Deep Learning for the Identification of Sudden Stratospheric Warming Events

曾怡甄
YI-JHEN Zeng

指導教授：梁禹喬 博士
Advisor: YU-CHIAO Liang, Ph.D.

中華民國 112 年 8 月

August 2023

doi:10.6342/NTU202303879

致謝

付梓之際，回首來路，兩年碩士生涯恍若隔日而已。

謹以此文獻給過去的自己。曾於大學時期經歷了一段困頓無光的時期，而在研究所的這些歲月，我從零開始整裝待發。自尋找問題至建立模型、自決定主題至最終完稿，期間不僅解決問題的能力快速增長、發現問題的眼光愈加銳利，我更好好地認識了自己，且比之過去更加從容與堅毅。

本研究的完成離不開許多人的幫助和支持，在此表達謝忱。

感謝我的指導教授，梁禹喬老師，從選題開始就給了我很大的空間，願意讓我去做各種嘗試、放我去學習各種我想要學的。在研究過程中，我能感受到我有充分的自主性，謝謝你相信我能夠獨立做好研究，你的信任與引導是我能從過去的低谷爬起來的非常重要的養分。

感謝金德學長協助進行二維訓練資料的 regridding、感謝王逸對訓練資料的整理及分類、感謝逸昌學長在建構類神經網路與訓練機器學習模型上的諸多寶貴與實用的建議。

感謝 Stack Overflow 和 GitHub 等平台，使我跨越了許多 coding 及建立模型時的瓶頸。

感謝我的論文口試委員羅敏輝老師、吳健銘老師及陳柏孚老師對這篇研究所給予的補充、建議、指正，當然還有肯定，整個碩士論文口試的過程我感到從容與放鬆，並且對未來可進行的方向有更多的想法。

最後，謝謝把我從困頓的低谷中一步步拉出來不曾放棄過的自己。

2023.08.08 於台北

doi:10.6342/NTU202303879

中文摘要

平流層爆發性增溫（Sudden stratospheric warming, SSW）事件是極地平流層中極端的天氣現象，在此一事件發生時，平流層極地渦旋的環流被破壞甚至逆轉，進一步在次季節的時間尺度上對地表產生後續的影響，了解平流層中的變異性與其和對流層之間的耦合關係，對於改進次季節至季節時間尺度上近地表場的預測十分重要。借助人工智慧在識別圖形及網格化資料上空間細節的潛力，我們可以藉由辨識 SSW 事件的任務來分析事件中平流層極地渦旋的空間結構。在本研究中，我使用全球氣候模式的系集模擬結果訓練機器學習模型辨識 SSW 事件，並採用了一種可解釋的深度學習方法測試機器學習的成效。我首先運用平流層位於 60°N 的一維緯向風場來訓練三種不同複雜度的類神經網路，從簡單到複雜依次為：邏輯回歸網路、淺層神經網路及深度神經網路。所有類神經網路均能以相當高的準確性識別 SSW 事件。為了瞭解這些類神經網路是如何學習區分 SSW 事件和非 SSW 事件，在測試階段，我沿著經度以不同長度的遮罩屏蔽掉一部分的緯向風場，以測試屏蔽一部分的資訊會如何影響類神經網路的表現，即藉由類神經網路對不同空間上的風場資訊的依賴性，來解析 SSW 的關鍵結構。當遮罩窗口較短時，淺層和深度神經網路並未表現出明顯的空間依賴性，而邏輯回歸網路則在約 160°W 附近表現出較強的空間依賴性，這一地區恰對應到緯向風的平均值為負、標準差較小之所在，隨著遮罩長度增加，淺層和深度神經網路逐漸顯現出空間依賴性。

為了進一步探討二維空間上的空間依賴性，我運用北半球以北極為中心的平流層二維緯向風場來訓練卷積神經網路，並透過不同大小的矩形遮罩來進行類似的測試。卷積神經網路對於二維風場的空間依賴性與前述一維的結果具一致性：空間依賴性較強的地方分布在負緯向風所在的區域，惟空間範圍擴展到 60°N 的北部和南部。除了緯向風場之外，我另使用了二維位勢高度場訓練卷積神經網路並進行相同的依賴性測試。卷積神經網路對於二維位勢高度場的空間依賴性可以分為兩個區域：1）阿拉斯加上空，對應到平均緯向風場負值所在區域，2）以北極為中心的圓形區域，前者與僅使用二維緯向風場訓練得到的結果相符，亦反映了風

doi:10.6342/NTU202303879

場和位勢高度場之間的地轉風平衡，而後者恰對應到 SSW 和非 SSW 事件之間的地形高度差異最大的地區。因此，卷積神經網路不僅能夠利用地轉關係，還學習到運用 SSW 和非 SSW 之間的對比，來區分 SSW 事件和非 SSW 事件。此研究的發現不僅表現了負緯向風場是 SSW 極地渦旋的關鍵特徵，也揭示 SSW 事件和 Northern Annular Mode (NAM) 負相位之間的關聯——NAM 負相位的特徵之一為極區壓力高於正常水平——。此外，額外的測試當中還顯示了 55°N 處的緯向風場也是辨識 SSW 事件的一個重要特徵，此發現補充了常見的 SSW 定義中所依據的 60°N 的重要性。

本研究的結果凸顯了可解釋性深度學習工具於學習 SSW 空間訊息和攫取關鍵性特徵的能力，並對於 SSW 肇始的追溯和隨後地表影響的預測可能具有重要意義。

關鍵字：可解釋性機器學習；平流層爆發性增溫事件；平流層極地渦旋；類神經網路；卷積神經網路

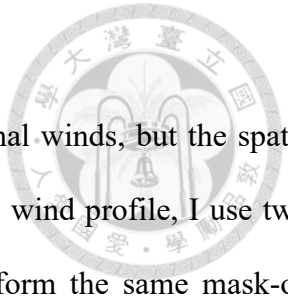
doi:10.6342/NTU202303879

ABSTRACT

An advanced understanding of stratospheric variability and its coupling to the troposphere is critical to improving the prediction of near-surface fields at subseasonal-to-seasonal timescale. In the most extreme situation, a sudden stratospheric warming (SSW) event occurs and substantially perturbs the stratospheric circulation and could, subsequently, exert profound surface impacts. Artificial intelligence could be a powerful tool in recognizing SSW spatial details resulting in a better categorization of the types of disrupted vortices. Here I apply an interpretable deep learning approach to identify SSW events from non-SSW ones using a large-ensemble suite of outcomes from a global climate model. I start with a one-dimensional case by using the stratospheric zonal wind profile of SSW events circling the 60°N latitude to train neural networks with different complexity: logistic regression network, shallow neural network, and deep neural network. All neural networks can identify SSW events with fairly high accuracy. To address the interpretability of how these neural networks learn to distinguish SSW from non-SSW events, I mask out the zonal wind fields with longitudinal windows with varying lengths to test if the spatial structure of disrupted winds is decisive for the network learning. Neither shallow nor deep neural networks show apparent spatial dependence when the masking window is short, while the logistic regression network gives stronger spatial dependence centering around 160°W, where small variation and negative mean value of zonal wind resides. The dependence of shallow and deep networks emerges as the window length increases.

To further explore the two-dimensional spatial dependence, I train a convolutional neural network (CNN) exploiting the two-dimensional zonal wind fields in the Northern Hemisphere. Similar tests are performed by strategically masking out the zonal wind fields by a rectangular region with varying sizes. The spatial dependence of two-dimensional neural network is largely consistent

doi:10.6342/NTU202303879



with one-dimensional networks, highlighting the region with negative zonal winds, but the spatial extents expand wider to the north and south of 60°N. In addition to zonal wind profile, I use two-dimensional geopotential height fields at 10 hPa to train CNN and perform the same mask-out analysis. Two regions of high spatial dependence emerge immediately: 1) the region where large negative zonal wind fields locate, and 2) a circular patch centering at the North Pole. The former corresponds well to the results obtained from the training using zonal wind profile solely, revealing the geostrophic wind balance between wind and geopotential height fields, whereas the latter is mainly associated with the largest geopotential height difference between SSW and non-SSW events. The CNN model, thus, utilizes not only the geostrophic relationship but also the large contrast between SSW and non-SSW, to learn the categorization task. These findings not only suggest the significance of negative zonal wind speed in characterizing a SSW polar vortex, as embodied in the common definition of SSW, but also unveil the association between SSW events and the negative phase of the Northern Annular Mode, which is characterized by the higher-than-normal geopotential height over the polar regions. Furthermore, additional tests revealed that the zonal wind field at 55°N is also an important feature in identifying SSW events, better capturing the core of negative zonal wind region and complementing the significance of 60°N on which common SSW definitions rely.

The above results highlight the capability of interpretable deep learning tools in learning the SSW spatial information and revealing the spatial dependence, which may carry out important implications for the prediction of SSW genesis and subsequent surface impacts.

Keywords: Interpretable deep learning; Sudden stratospheric warming events; Stratospheric polar vortex; Neural networks; Convolutional neural network

doi:10.6342/NTU202303879

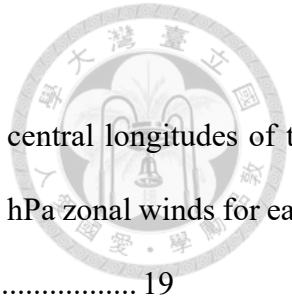
CONTENTS

doi:10.6342/NTU202303879

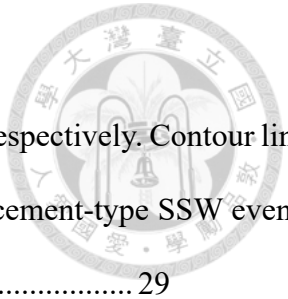
LIST OF FIGURES

doi:10.6342/NTU202303879

doi:10.6342/NTU202303879



doi:10.6342/NTU202303879



doi:10.6342/NTU202303879

Chapter 1 Introduction

Sudden Stratospheric Warmings (SSWs) are the primary dynamical events observed and simulated in the wintertime stratosphere of the Northern Hemisphere (Baldwin et al. 2021). When a major SSW event occurs, the stratospheric temperature can rise 30-40K in a few days accompanied by the apparent distorted structure of polar vortex (Limpasuvan et al. 2004), leading to, in the zonal-mean sense, the stratospheric westerly reversal (Charlton and Polvani 2007). Progress has been made in developing theoretical framework to understand the driving mechanisms of SSWs, which mainly involve the interaction between the zonal flow of the polar stratosphere and upward propagation of planetary waves (Charney & Drazin 1961; Matsuno 1971; Schoeberl 1978; Andrews et al. 1987; Garfinkel et al. 2012). Some studies also showed that SSWs can arise from tropospheric precursors, including the upward wave activity fluxes (Polvani & Waugh 2004), the blocking system in North America, Greenland, and Eurasia (Martius et al. 2009; Nishii et al. 2011; Barriopedro & Calvo 2014), or a sea-level pressure dipole in the subtropical-subpolar North Pacific (Cohen & Jones 2011; Lehtonen & Karpechko 2016; Dai & Hitchcock 2021; Dai et al. 2023). SSWs can also be triggered solely by the stratospheric internal dynamics (Scott & Polvani 2004; 2006; Hitchcock & Haynes 2016; de La Cámara et al. 2019).

The surface impacts of SSWs, via the downward influence, have been generally described as a negative phase of the North Atlantic Oscillation (NAO) (e.g., Baldwin & Dunkerton 1999, 2001; Charlton & Polvani 2007; Kolstad et al. 2010; Sigmond et al. 2013; Hitchcock & Simpson 2014; Kidston et al. 2015; Butler et al. 2017; Hall et al. 2021). The pronounced features include a southward shift of Atlantic jet together with reduced

doi:10.6342/NTU202303879

stormtrack activities, and a temperature decrease over northern Europe and Siberia (Kidston et al. 2015; Domeisen et al. 2020; Dai & Hitchcock 2021; Dai et al. 2023). In particular, a subset of SSWs, referred to as Polar-night Jet Oscillation events, presents a very persistent warm signal in lower stratosphere that is accompanied by strong and persistent tropospheric impacts (Hitchcock et al. 2013). Recent studies further highlighted the role of stratosphere-troposphere-ocean coupling in the North Pacific in modulating the spatial distribution of SSW surface influences (Dai & Hitchcock 2021; Dai et al. 2023). More importantly, the downward influences of SSWs, characterized by a timescale from a few weeks up to two months on the tropospheric circulation, could benefit the exploration of subseasonal-to-seasonal predictability sources for tropospheric variables (Baldwin et al. 2003; Charlton & Polvani 2007; Sigmond et al. 2013; Kidston et al. 2015; Davis et al. 2022).

Recent progress in artificial intelligence (AI) has spurred considerable interests in its applications in a wide spectrum of earth system science problems (e.g., Reichstein et al. 2019; Li et al. 2023). Despite several data-driven methods were used to enhance our understanding of SSW dynamics and explore the potential prediction skill for SSW genesis, including principal component analysis/empirical orthogonal functions (Ren & Cai 2006; Lu & Ding 2015; Lu et al. 2016; de Fondeville et al. 2023), vortex geometry (Hannachi et al. 2011; Mitchell et al. 2011), causal discovery algorithm (Kretschmer et al. 2017), computer vision (Lawrence & Manney 2018), deep learning studies on the SSWs, however, they were not widely appreciated mainly due to the limited number of events in observational records and reanalysis datasets (de Fondeville et al. 2023). Indeed, only 35 to 38 SSWs during the 1958-2014 period have been identified in the past 6 decades in observational records (Butler et al. 2017), which is far way too small for deep learning models to be trained. Instead of targeting SSWs per se, previous studies trained

doi:10.6342/NTU202303879

their machine or deep learning models using stratospheric and tropospheric daily (or hourly) fields from reanalysis datasets and examined the characteristics, evolutions, and predictors of SSWs as extreme events in the stratosphere (Blume et al. 2012; Wu et al. 2022; de Fondeville et al. 2023). However, no training exercise using SSW events directly, which should lead to better characterization of SSWs, has not been carried out and documented according to my literature survey. This motivates us to apply deep learning techniques to examine SSWs directly, potentially providing a new perspective on their spatial and temporal characteristics.

As deep learning tools have become popular and useful in many earth system problems, the perception that neural networks are “black boxes” needs to be addressed since their decision-making processes are incomprehensible (Savage 2020). Recent studies have emphasized the significance of interpretability in machine learning models from different perspectives (Barnes 2022). For example, DeGrave et al. (2021) demonstrated that medical imaging AI neglects important features, causes a decision-making basis, and leads to increased likelihood of incorrect diagnoses. Some interpretability methods have been implemented to decipher how machine learning tools learn and to discover new science (Barnes 2020). Toms et al. (2020), for example, presented the ability of two interpretable machine learning methods in exposing the physically meaningful patterns in ENSO prediction problem. As such, this study aims to enhance the interpretation, rather than focusing on the classification performance per se, to emphasize how neuronal networks learns or utilize information to make decisions.

The structure of this paper is outlined as follows. Chapter 2 provides an overview of the data used in this study and the definition of SSW events, followed by descriptions of the structure of the neural networks used and the interpretability method applied. The novelty of this study is that the deep learning models directly use SSW events, which has

doi:10.6342/NTU202303879

not been attempted in previous studies according to my literature survey. Chapter 3 presents the results obtained from the tests using the interpretability method. Chapter 4 provides the discussion and Chapter 5 presents the conclusion.

doi:10.6342/NTU202303879

Chapter 2 Data and Methods

The number of SSW events observed since 1958 is no more than 40 (Butlet et al. 2015), which is insufficient for training a deep learning model. To obtain greater amount of training data, I exploit the output of a suite of large-ensemble simulations using the Whole Atmosphere Community Climate Model Version 6 (WACCM6) (Gettelman et al. 2019). Following the protocol developed by the Blue-Action Project (https://blue-action.eu/index.php?id), global daily ¼ degree sea surface temperatures (SSTs) and sea ice concentrations (SICs) during the 1979–2014 period from U.K. Met Office Hadley Centre Sea Ice and SST Version 2.2.0.0 (Kennedy et al. 2017; Rayner et al. 2003; Titchner & Rayner 2014) were used to force WACCM6. The first set of experiment (hereafter EXP1) was prescribed the time-varying SST and SIC fields, while the second set (hereafter EXP2) replaces the SIC field of the Northern Hemisphere by its climatological (1979-2014) values. Each experiment contains 30 members in order to examine the role and effect of internal variability.

To identify the SSW events, I, following the method of Charlton & Polvani 2007 (hereafter CP07) to define the first date of SSW events when the zonal-mean zonal winds (hereafter U) at 60°N and 10 hPa is less than 0 (i.e., the zonal wind becomes easterly) as the central date of an SSW event. In this way, I obtain 759 SSW events in EXP1 and 780 events in EXP2. Data for non-SSW events are extracted from climatological data by a random sampling approach with the daily zonal-mean zonal winds at 60°N, 10 hPa is greater than 0 (i.e., westerly). For the sake of computation efficiency, I reduce the spatial resolution to 2.5° x 2.5°. Although the SIC forcing is different in EXP1 and EXP2, the characteristic and statistics of SSW does not show substantial differences (Wang et al.

doi:10.6342/NTU202303879

2022, in preparation). This resonates with the findings of Liang et al. (2019), using the same dataset, that SIC impacts on large-scale atmospheric circulation are rather small and limited in local area where sea ice experiences substantial retreat. **Figure 1**a shows the 10 hPa geopotential height (GPH) and zonal wind (U) fields for a non-SSW event, while **Figure 1**b and **Figure 1**c are the displacement and splitting types of SSW events, respectively in EXP1.

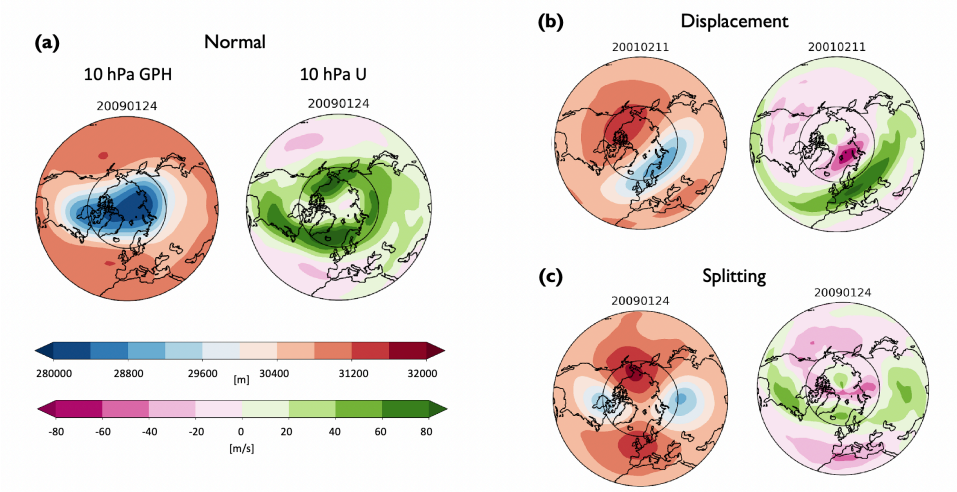


Figure 1: 10 hPa geopotential height (GPH) and zonal wind (U) fields of (a) normal (non-SSW), (b) displacement and (c) splitting type SSW events from WACCM6 EXP1. The 60°N line is marked as a black circle on each map, delineating the latitude chosen in CP07.

2.1 One-dimensional Case

In the one-dimensional (1-D) case, only 10 hPa U circling 60°N from EXP1 is used to train the neural networks. The Python package Pytorch (Paszke et al. 2017) is used to build all the neural network models used in this study. To compare the performance of

doi:10.6342/NTU202303879

models with different complexity, I build three neural networks, from simple to complex, namely logistic regression network, shallow neural network (SNN), and deep neural network (DNN). **Figure 2** illustrates the architecture of the three neural networks used for 1-D SSW identification task. Specifically, the logistic regression network consists of an input layer of 144 neurons, corresponding to U at 10 hPa of 360° longitude at 2.5° spatial resolution, and an output layer of a single neuron indicating whether it is an SSW event or not. The SNN comprises a 144-neuron input layer, a 30-neuron hidden layer, and a single-neuron output layer. The structure of DNN is the same as SNN except that DNN has one more hidden layer of 30 neurons. In all models, the last layers before the output layers incorporate the sigmoid function as the activation function, while the rest parts of the neural networks use ReLU. All models use the binary cross entropy loss function. The stochastic gradient descent is used as the optimization algorithm.

During the training process, 759 SSW and non-SSW events from EXP1 are used, 70% of which are used as the training set and the rest as the validation set. In the testing stage, SSW and non-SSW events from EXP2 are used to assess the error and interpretability based on whether or not the trained neural networks can distinguish between SSW events and non-SSW ones. To enhance robustness, I randomly draw 200 samples from 780 SSW/non-SSW events per test and repeat 50 times, giving a total of 10,000 outcomes for evaluating the performance and interpretability.

doi:10.6342/NTU202303879

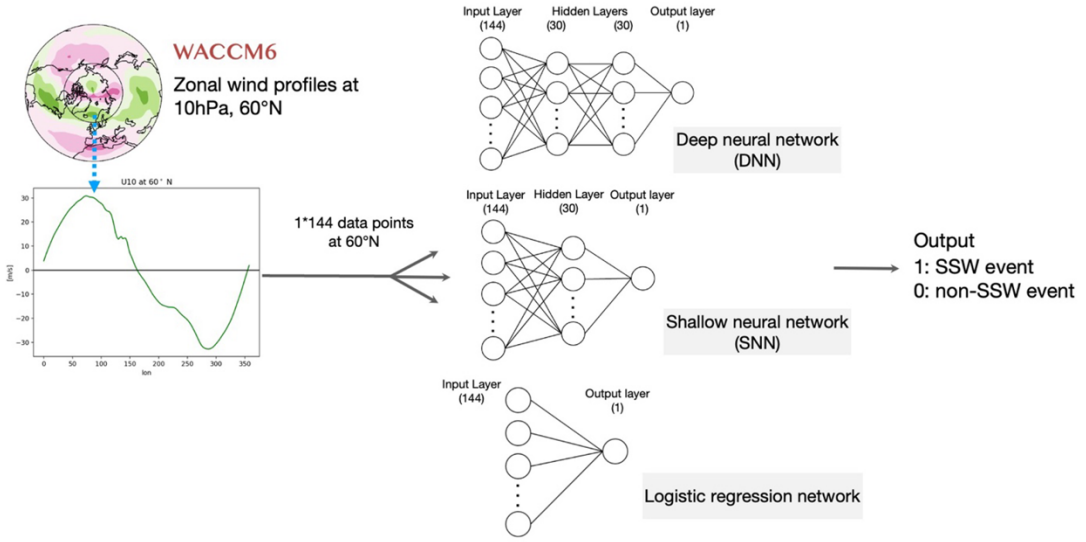


Figure 2: The architecture of the models used for 1-dimensional SSW data classification

To address interpretability, I mask out input zonal wind fields with varying longitudinal windows (30° to 180° longitudinal length) by setting their values to 0. This allows us to assess the impact of removing particular regions on the overall model performance. The percentages of misclassifications out of 10,000 testing versus the central longitude of the masking are shown in **Figure 4** and **Figure 5**.

2.2 Two-dimensional Case

The two-dimensional (2-D) U and geopotential height (GPH) fields at 10 hPa in the Northern Hemisphere are used for training and testing in the two-dimensional case. To cope with the potential influence of the uneven distribution of data points in the WACCM6 output, which follows the Gaussian grid with an increased grid point density

doi:10.6342/NTU202303879

near the poles, I employ a bilinear interpolation to interpolate the U and GPH data onto the cubed-sphere grid (Jung et al. 2019). The training and testing data encompass the area approximately north of 45°N, which corresponds to the top panel of the cubed sphere, comprising 36x36 grid points.

Here I use a convolutional neural network (CNN) with an architecture as follows:

$$[INPUT] \rightarrow [CONV1] \rightarrow [BATCH\ NORM] \rightarrow [ReLU]$$
$$\rightarrow [CONV2] \rightarrow [BATCH\ NORM] \rightarrow [ReLU] \rightarrow [POOL]$$
$$\rightarrow [CONV3] \rightarrow [BATCH\ NORM] \rightarrow [ReLU]$$
$$\rightarrow [CONV4] \rightarrow [BATCH\ NORM] \rightarrow [ReLU]$$
$$\rightarrow [FC\ Layer] \rightarrow [OUTPUT]$$

Figure 3 illustrates the architecture of the CNN. The network consists of a total of 8 layers. The input image with a dimension of 36×36 undergoes convolution in two consecutive convolutional layers. Following this, the third layer performs a pooling operation, which is then followed by two convolutional layers. In each of the convolutional layers, I use a kernel of spatial size 3×3 with stride size of 1 and padding of 1. The pooling layer employs max pooling with a kernel size of 2, stride of 2, and zero padding. Once the dimension is reduced, the sixth layer is a fully connected layer, with 5,184 neurons flattening the feature map. The seventh layer is another fully connected layer with 10 neurons. The final eighth layer is a single-neuron sigmoid output layer indicating the SSW/non-SSW event. All except for the last layer uses ReLU as the activation function.

doi:10.6342/NTU202303879

Same as 1-D case, the training process involves utilizing 759 SSW and non-SSW events from EXP1, with 70% of the events allocated as the training set and the rest as the validation set. U and GPH fields are used to train CNN separately. In the testing stage,

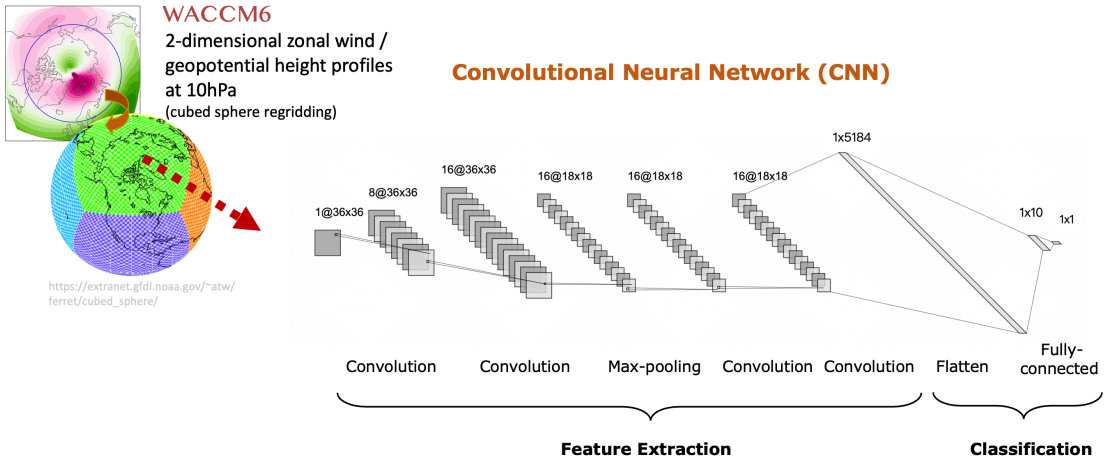


Figure 3: The architecture of the convolutional neural network used for 2-dimensional SSW data classification. The input layer is 2-D U or GPH field regridded onto cubed-sphere grid, covering the area north of 45°N with 36x36 grid points.

SSW and non-SSW events from EXP2 are used to calculate the error. Similarly, I perform mask-out tests in which rectangular regions of different sizes (3×3, 5×5, 7×7, 9×9, 11×11, and 13×13 grids or equivalently 6.2°×6.2°, 10.3°×10.3°, 14.5°×14.5°, 18.6°×18.6°, 22.8°×22.8°, and 26.9°×26.9°) are employed to systematically remove a rectangular portion of the U and GPH fields by setting their values to 0. The misclassification rates of U and GPH fields with the central grids of the rectangular masks are shown in **Figures 6**, **7** (U) and **8**, **9** (GPH).

doi:10.6342/NTU202303879

Chapter 3 Results

I start with giving the unmasked 1-D U data at 60°N, 10 hPa, to train the SNN and DNN, which perform very well on this SSW identification task with nearly zero error rates. In contrast, the logistic regression network exhibits misclassifications, identifying SSW as non-SSW 121 times (and non-SSW as SSW 0 times) out of 20,000 tests.

Figure 4 presents the identification results of masked 1-D SSW U data, with each panel labeled as (a)-(f) corresponding to mask sizes of 30°, 60°, 90°, 120°, 150°, and 180°, respectively. The results are expressed in the form of percentages representing the misclassification of SSW events as non-SSW, and are plotted against the central longitudes of the masks. Overall, SNN and DNN perform much better than logistic regression network. When the masked longitudinal window is smaller, neither SNN nor DNN shows apparent spatial dependence, while the logistic regression network gives strong spatial dependence centering around 160°W (**Figure 4** (a)-(c)). As the masked range increases, both SNN and DNN exhibit slight spatial dependencies around 160°W, consistent with the logistic regression network and where small variation and negative mean values of zonal wind reside. The mean and standard deviation of 60°N, 10 hPa U of 780 EXP2 SSW events are shown in gray solid and dashed lines, respectively. The locations where the maximum error percentage resides align better with the trough of negative mean value than that of standard deviation.

One might immediately ask about the identification results of masked non-SSW events. The misclassification rates of 1-D non-SSW U versus the central longitudes of the masks are shown in **Figure 5** (a)-(f). Grey solid and dashed lines represent the mean and standard deviation of 60°N, 10 hPa zonal winds of 780 EXP2 non-SSW events. All three

doi:10.6342/NTU202303879

1-D neural networks are unlikely to misidentify non-SSW events as the errors are negligible. The results provide a clear indication of what neural networks learn to distinguish SSW from non-SSW events. The areas where the logistic regression network performs poorly align with the regions characterized by negative mean U values, indicating that negative zonal wind values (i.e., easterlies) play a crucial role in distinguishing SSW events from non-SSW ones, which was also the key field CP07 highlighted. Both SNN and DNN capture the characteristics of SSW vortices and thus do not show prominent spatial dependencies. To ensure the reliability of the findings, I conduct two additional experiments: one using EXP2 data for training and EXP1 for testing, and the other using a mixture of EXP1 and EXP2 data for training and testing. The results of both additional experiments closely resemble those illustrated in **Figure 4** and **Figure 5**. (APPENDIX **Figure 12** and **Figure 13**)

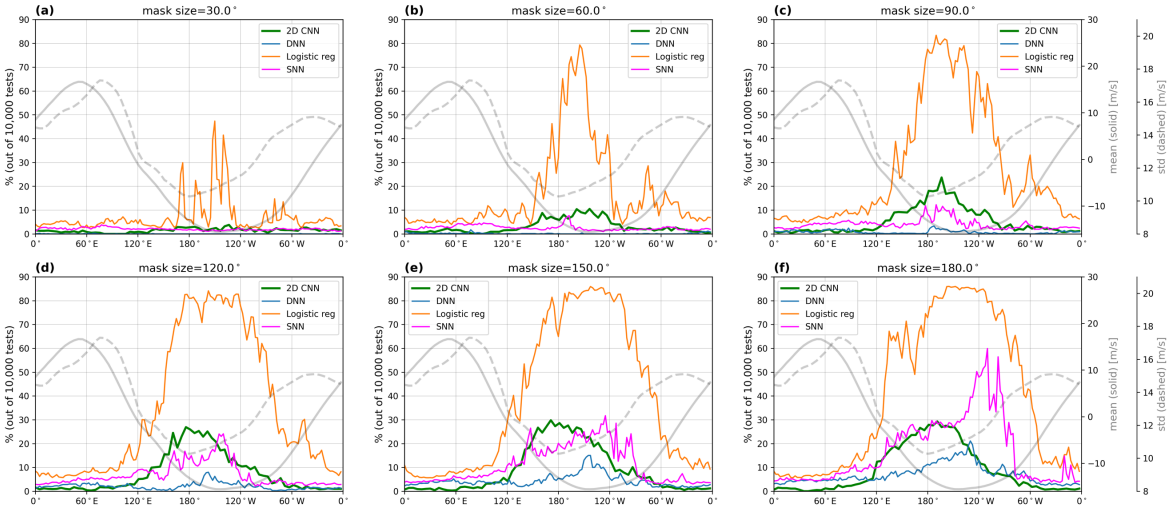


Figure 4: The error SSW classification rates of logistic regression network, SNN and DNN as a function of the central longitudes of the masks. (a)-(f) represent the masked longitudinal window of 30°, 60°, 90°, 120°, 150°, 180°, respectively. The mean (solid) and standard deviation (dashed) of 60°N, 10 hPa zonal winds of 780 EXP2 SSW events are also shown.

doi:10.6342/NTU202303879

Considering the potential bias in WACCM6 model simulation, I perform testing by inputting reanalysis data into the models trained with EXP1. The reanalysis data used are products from NCEP/NCAR Reanalysis Project (Kalnay et al. 2018), containing 22 SSW events from 1980 to 2014. All three 1-D networks successfully classify all 22 SSW events. The misclassification count of the 22 SSW events based on 1-D U with the central longitude of the masks is illustrated in APPENDIX **Figure 14**. The spatial distribution of the error counts closely aligns with the results in **Figure 4**, indicating the models trained with WACCM6 large-ensemble simulation data adeptedly capture the key spatial characteristics of actual SSW events.

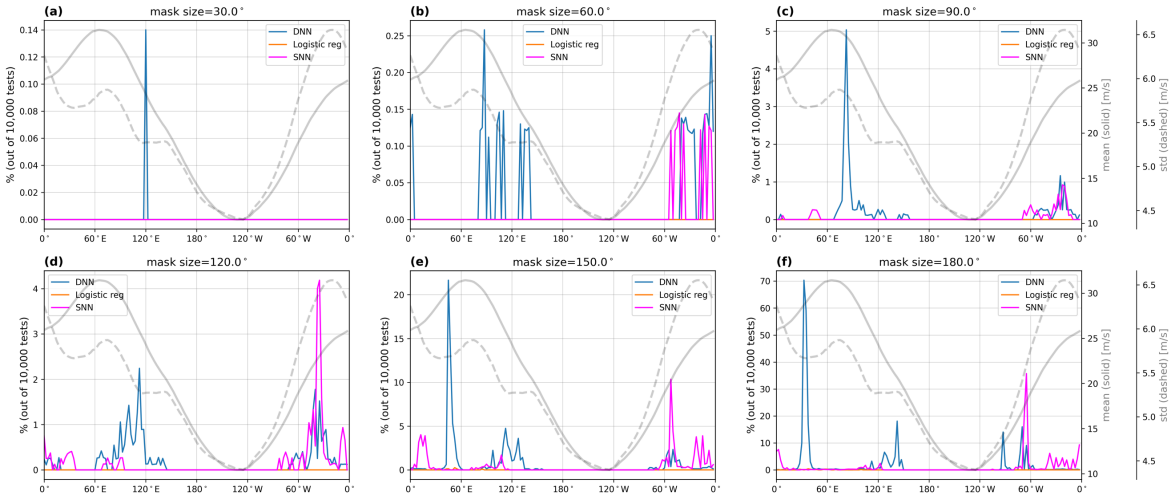


Figure 5: The error non-SSW classification rates of logistic regression network, SNN and DNN as a function of the central longitudes of the masks. (a)-(f) represent the masked longitudinal window of 30°, 60°, 90°, 120°, 150°, 180°, respectively. The mean (solid) and standard deviation (dashed) of 60°N, 10 hPa zonal winds of 780 EXP2 non-SSW events are also shown.

To further enhance the understanding of the dependence of deep learning models on 2-D spatial structure, my next step is to train a CNN which explicitly exploits 2-D U fields

doi:10.6342/NTU202303879

and assess its performance and interpretability. **Figure 6** illustrates the SSW misclassification rates based on the central grid of the rectangular masks. Each panel, labeled as (a)-(f), corresponds to a rectangular mask with the box widths/lengths of 6.2°, 10.3°, 14.5°, 18.6°, 22.8°, and 26.9°, respectively. The contour lines represent the mean 10 hPa U fields of 780 SSW events. The blue lines mark the position of 60°N for reference. As the masked area increases, the CNN demonstrates significant spatial dependence primarily above Alaska and northern Canada, overlapping the region with negative mean

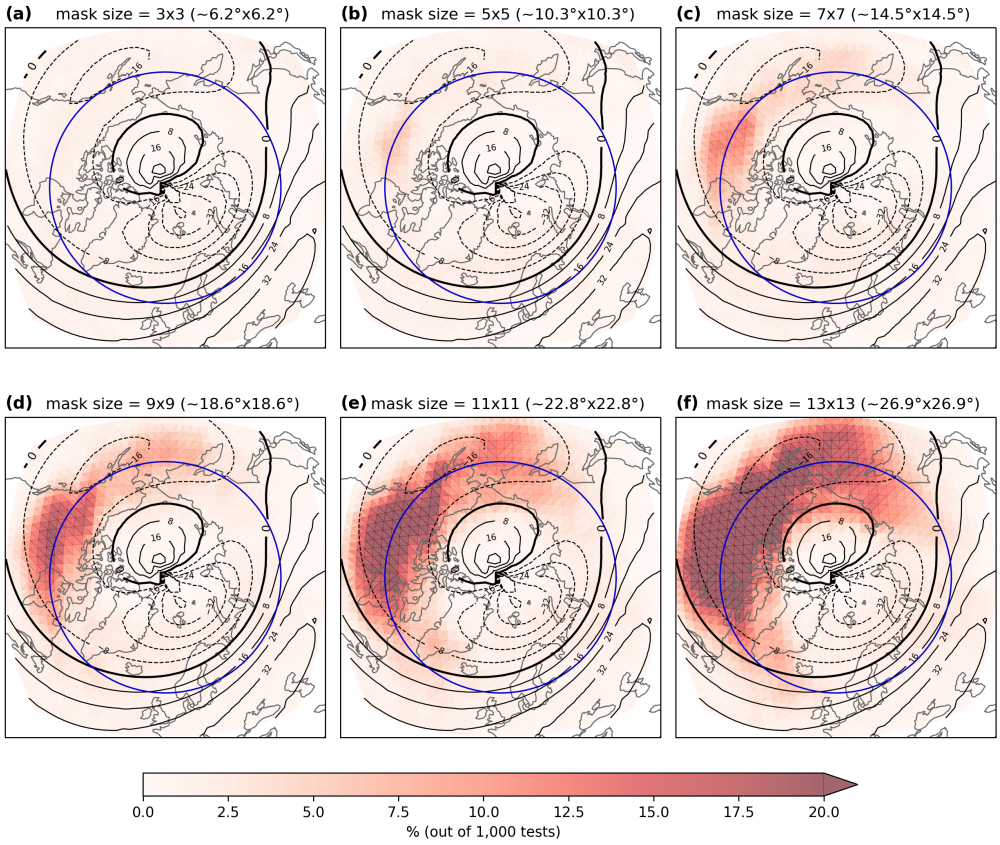


Figure 6: Red patches: the error SSW classification rates of zonal wind fields with the central grids of the rectangular masks. (a)-(f) represent rectangular masking regions with side length of 6.2°, 10.3°, 14.5°, 18.6°, and 22.8°, respectively. Contour lines: the mean 10 hPa zonal wind field (in m/s) of 780 EXP2 SSW events. The longitude 60°N is marked with a blue circle in each map.

doi:10.6342/NTU202303879

values of zonal winds over 60°N. This spatial dependence is largely consistent with that of the 1-D networks (**Figure 4**), but the spatial extents expand wider to the north and south of 60°N. Regarding the identification results of masked non-SSW, the areas where CNN demonstrates dependence are mainly distributed outside 60°N (**Figure 7**), agreeing with the findings in 1-D masked non-SSW identification.

There is a strong agreement between the results of the dependency tests conducted on 1-D and 2-D neural networks. In view of this and in order to extract more spatial

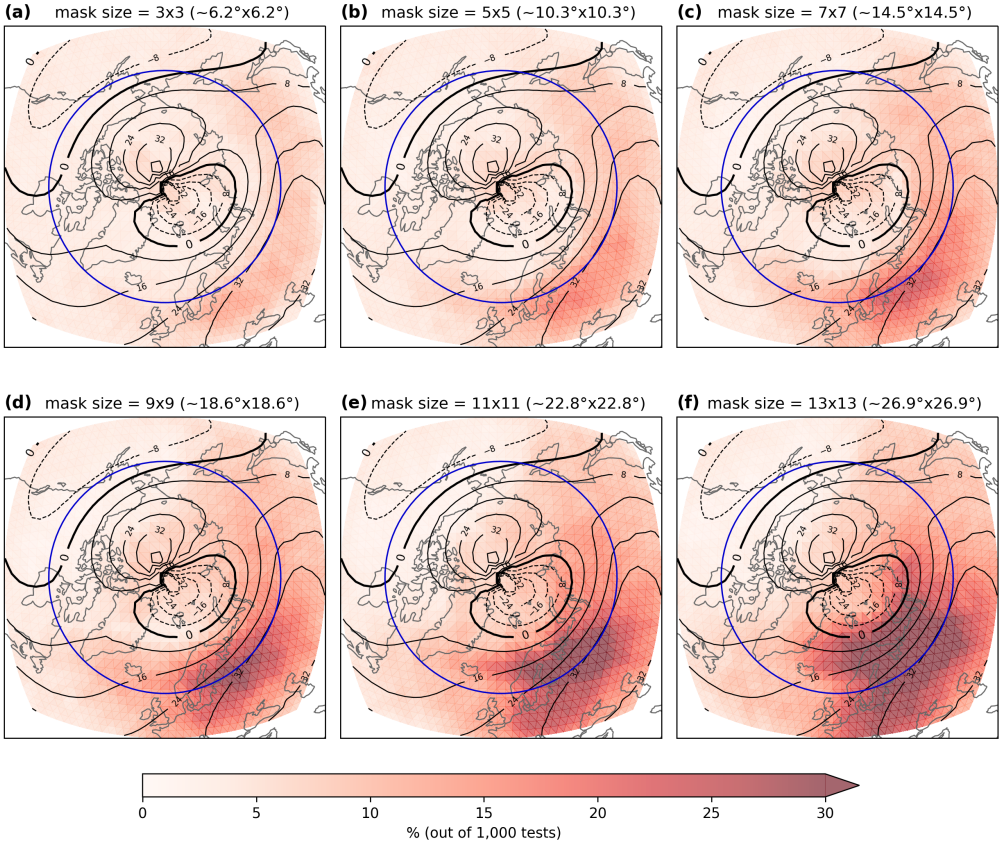


Figure 7: Red patches: the error non-SSW classification rates of zonal wind fields with the central grids of the rectangular masks. (a)-(f) represent rectangular masking regions with side length of 6.2°, 10.3°, 14.5°, 18.6°, and 22.8°, respectively. Contour lines: the mean 10 hPa zonal wind field (in m/s) of 780 EXP2 non-SSW events. The longitude 60°N is marked with a blue circle in each map.

doi:10.6342/NTU202303879

characteristics, I proceed to train the CNN using 2-D GPH fields and perform the spatial dependence tests to find the important spatial features of SSW vortices. **Figure 8** displays the SSW misclassification rates using rectangular masks of different sizes, with colored contour lines representing the mean 10 hPa GPH fields of 780 SSW events and blue lines indicating 60°N as a reference. The spatial dependence of the CNN trained using GPH fields can be divided into two key regions. The first region locates above Alaska, which corresponds to the key region identified using 2-D U fields, reflecting the negative zonal-

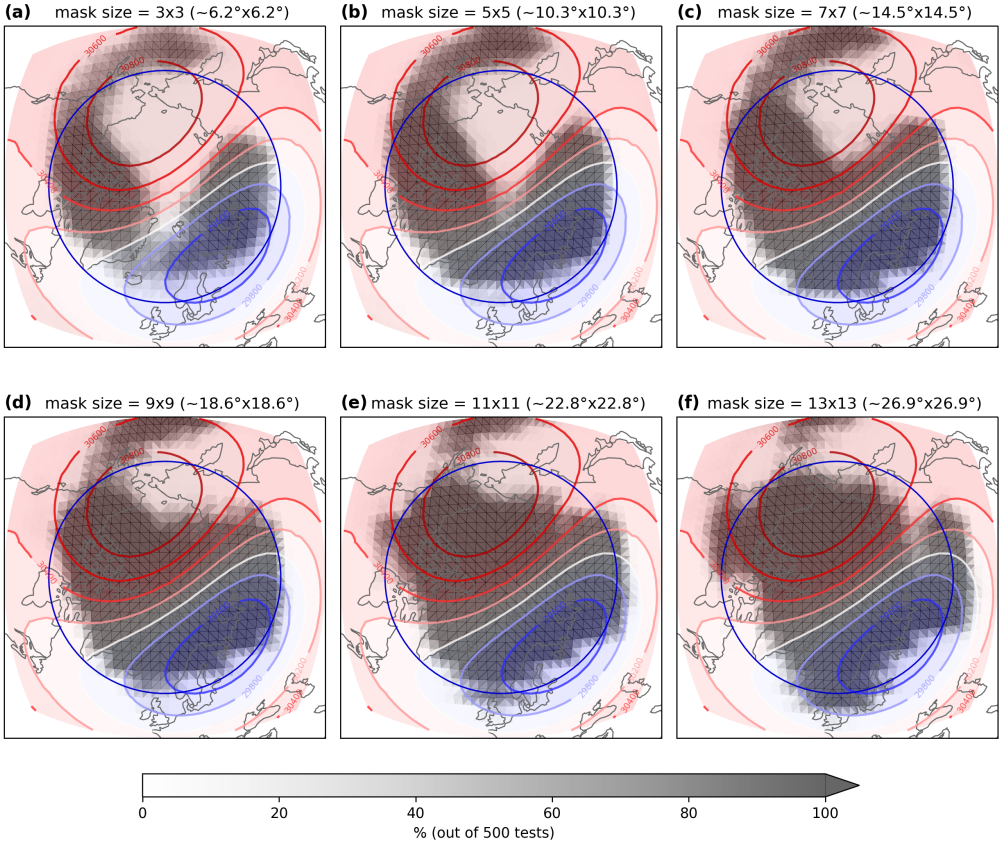


Figure 8: Gray patches: the error SSW classification rates of geopotential height fields with the central grids of the rectangular masks. (a)-(f) represent rectangular masking regions with side length of 6.2°, 10.3°, 14.5°, 18.6°, and 22.8°, respectively. Colored contour lines: the mean 10 hPa geopotential height field (in m/s) of 780 EXP2 SSW events. The longitude 60°N is marked with a blue circle in each map.

doi:10.6342/NTU202303879

wind field resides. The second region is centered around the North Pole, which corresponds to the area exhibiting the most significant difference between SSW and non-SSW mean GPH fields (**Figure 10**a) and the region where the standard deviation difference between SSW and non-SSW GPH fields is relatively smaller (**Figure 10**b). Such findings reveal that the CNN model learns the key structure of SSW through two aspects. The first region reinforces the importance of negative zonal wind values as shown in 2-D U fields. The second region points out that the CNN learns to capture the spatial

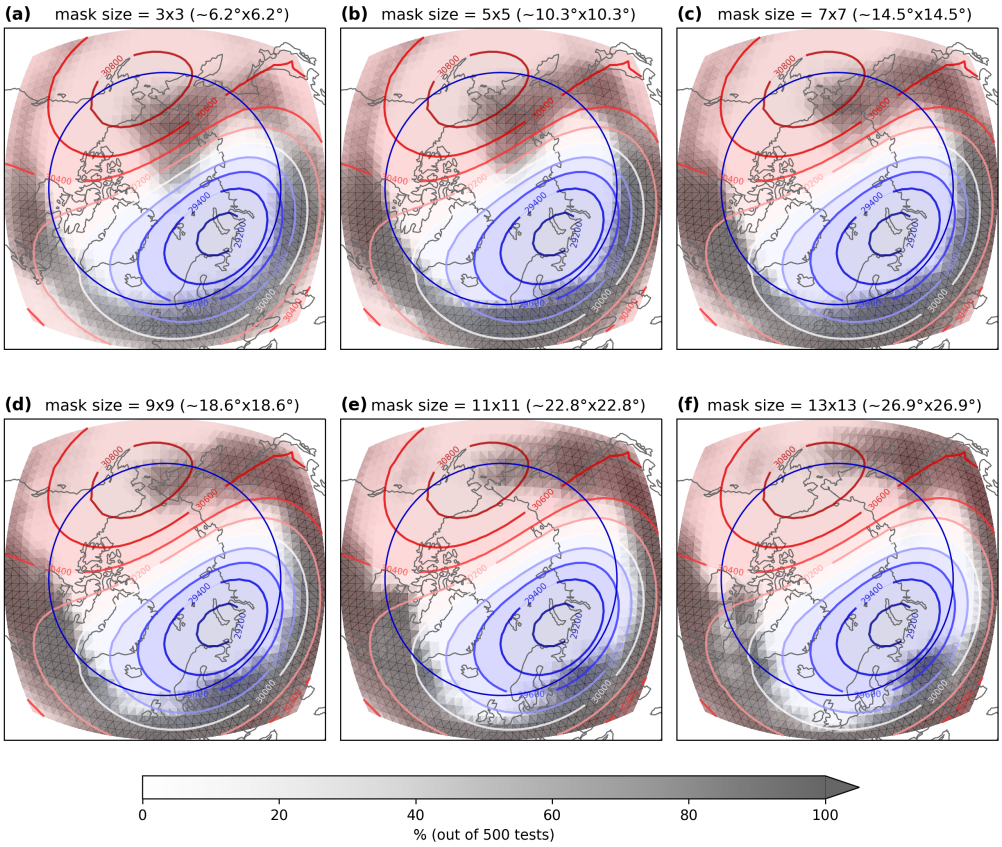


Figure 9: Gray patches: the error non-SSW classification rates of geopotential height fields with the central grids of the rectangular masks. (a)-(f) represent rectangular masking regions with side length of 6.2°, 10.3°, 14.5°, 18.6°, and 22.8°, respectively. Colored contour lines: the mean 10 hPa geopotential height field (in m/s) of 780 EXP2 non-SSW events. The longitude 60°N is marked with a blue circle in each map.

doi:10.6342/NTU202303879

contrast between SSW and non-SSW GPH fields and are thus able to distinguish them. The areas where the CNN exhibits dependencies in the identification of masked non-SSW GPH fields are primarily located outside of 60°N (**Figure 9**), which is consistent with the results derived from the identification of 2-D masked non-SSW U fields.

The results of the 2-D dependence test also reveal the significance of the zonal wind information at 60°N for recognizing SSW events. Thus, I examine the performance of CNN only when the training data around 60°N are masked out. To investigate this, I conduct dependence tests by specifically masking out the zonal wind field around 60°N with varying longitudinal windows (30°-180°). The results of these tests are presented in **Figure 4** as thick green lines.

Additionally, I extend the investigation by masking out the zonal wind fields around

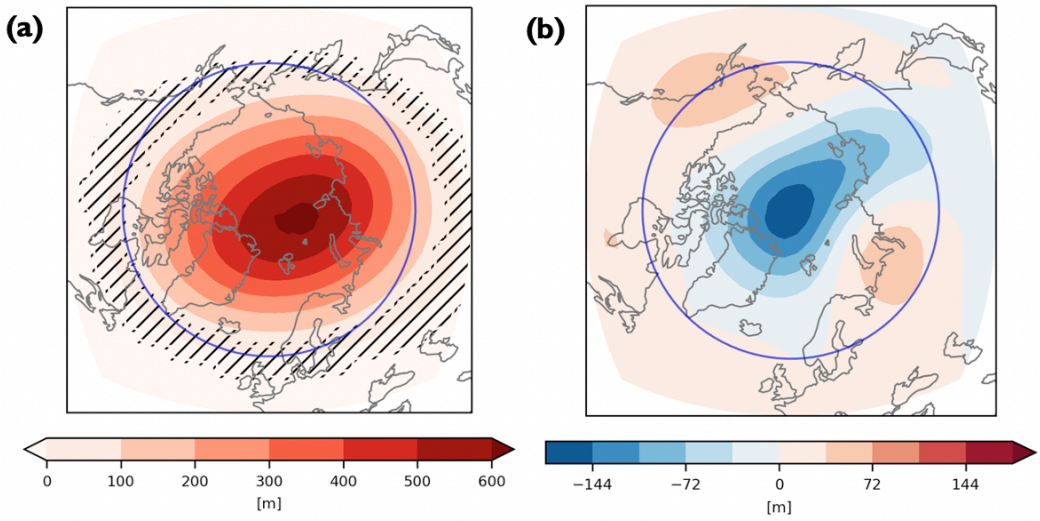


Figure 10: (a): The difference in GPH fields between SSW and non-SSW event (SSW – non-SSW). The hatched area represents the portion that does not pass the $p \leq 0.05$ significance test. (b): The standard deviation of SSW GPH fields minus the standard deviation of non-SSW GPH fields. The SSW and non-SSW data used here are all from EXP2.

doi:10.6342/NTU202303879

55°N, 65°N, and 75°N. The results of specifically masking out the zonal wind fields at these latitudes with a longitudinal window of 120° are compared in **Figure 11**. The strongest spatial dependence is observed around 160°W at 60°N. As the latitude reaches 65°N and 75°N, the spatial dependence decreases significantly. It is noteworthy that when the zonal wind centered around 160°W at 55°N is masked out, CNN still exhibits a certain degree of misclassification rate. The result suggests that, in addition to 60°N, which is the basis for defining SSW events in this study, the information around 55°N is also important for identifying SSW events.

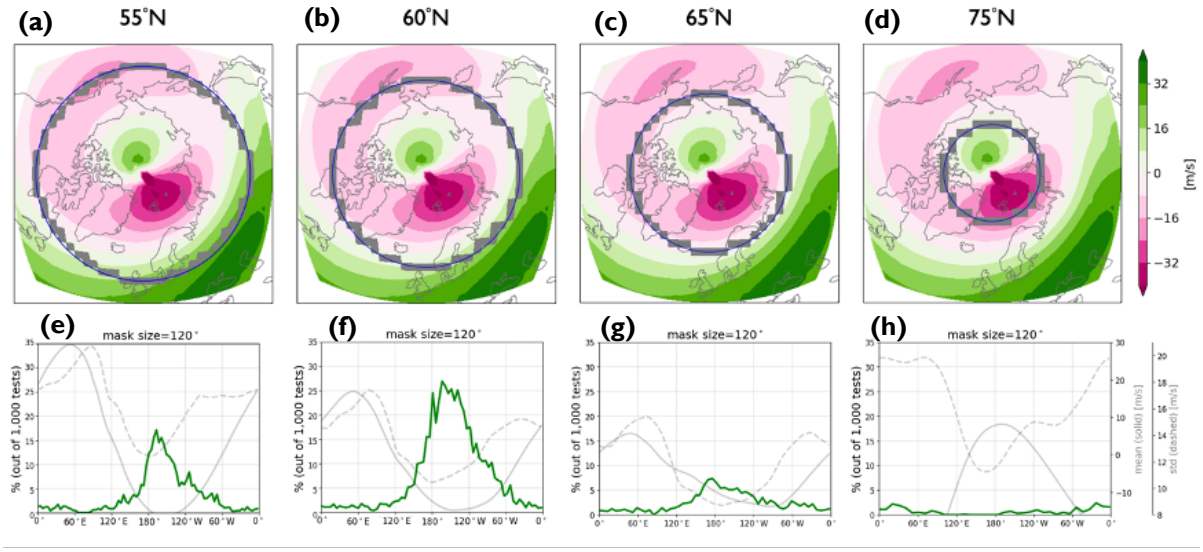


Figure 11: Each column corresponds to the masking of zonal winds around 55°N, 60°N, 65°N and 75°N. (a)-(d): The corresponding latitudes (blue circles) and the mask locations (gray patches) superimposed on the mean 10 hPa zonal wind field of 780 EXP2 SSW events. (e)-(h): The error SSW classification rates as functions of the central longitudes of the masks. The mean (solid) and standard deviation (dashed) the 10 hPa zonal winds for each latitudes of 780 EXP2 SSW events are also shown.

doi:10.6342/NTU202303879

Chapter 4 Discussion

Through interpretability deep learning approaches, this study unravels how deep learning models learn to effectively categorize SSW events. I find that deep learning models learn the SSW spatial structure utilizing the negative zonal wind values and the largest contrast between SSW and non-SSW geopotential height fields in terms of the mean and variation. I also find that apart from 60°N, being the basis for the definition of SSWs, the zonal winds at 55°N can also provide information for CNN to identify SSW events.

A follow-up question is whether or not the deep learning models can distinguish different types of SSW events. To investigate, I perform similar mask-out analysis for splitting and displacement SSWs, and find no substantial difference in the spatial dependencies between them (ref: APPENDIX **Figure 15** & **Figure 16**). This finding may reflect the internal variability of the two types of SSW, as the composites of displacement-type and splitting-type SSWs resemble each other. Training neural networks and assessing the spatial dependence of each type of event separately may help elucidate the different spatial features between them.

The results presented in this study could provide additional information for SSW genesis and may shed insights into the capability and process-based understanding of the deep learning model performance and interpretability. For example, a previous study found a link between SSW events and preceding tropospheric blocking precursors, both of which the blocking pattern and accompanying wave signal displayed significant distinction between displacement-type and splitting-type SSWs (Martius et al. 2009). Prior to displacement-type events, there exhibits a strong upward signal of zonal

doi:10.6342/NTU202303879

wavenumber1, while the wavenumber2 signal is nearly absent. For splitting-type events, both wavenumber1 and wavenumber2 signals are prominent. Decomposing the training data into wavenumbers 1 and 2 signals may help gain deeper insights into the key spatial features of SSW, which can be an extended study. Additionally, implementing the data from the period prior to the SSW onset dates, say day -10 to -1, to the training dataset may assist the deep learning models to better explore the key SSW spatial structure more accurately, as previous studies showed that information prior to the central date is important for SSW genesis and evolution (Polvani & Waugh 2004; Martius et al. 2009; Cohen & Jones 2011).

Exploiting neural networks to further understand the subsequent downward influence of SSW events is also an aspect that my future study can focus on. Such an attempt would require data with broader spatial and temporal coverage, including lower-level fields and surface variables from a few weeks up to two months after the SSW onset dates. In the training process. Correspondingly, the computational resources required for this analysis will also increase significantly.

To validate the results, other model interpretability methods (e.g., occlusion, integrated gradients, layerwise relevance propagation, etc.) can be utilized. In the interpretability approach adopted in this study, I perturb the input SSW features by setting the value of the masked-out regions to 0 to investigate how changes to the input features correspond to the error of the final model prediction. However, it is essential to note that such operations that set a value of 0 may affect the results, given that this study relies on the negative/positive values of zonal mean zonal wind to define and label SSW/non-SSW events. As an alternative approach for addressing this issue, instead of setting the value of the masked-out regions to 0, I could consider dropping out the input neurons corresponding to the masked area. But such an approach would change the architecture

doi:10.6342/NTU202303879

of the neural network itself, thereby raising the question of how varying the architecture of a neural network would affect the results.

A more objective method, such as the Shapley value (Shapley 1953), could be applied to my problem. Shapley values provide a way to quantify the marginal contribution of each feature (which corresponds to each grid of the SSW wind profile) to the prediction of a model. This is done by "excluding" features through sampling the empirical distribution of the feature's values and averaging over multiple samples (Gopinath 2021), and may offer valuable insights into the significance of each spatial feature of SSW vortices.

doi:10.6342/NTU202303879

Chapter 5 Conclusion

In this study, I have utilized an interpretable deep learning approach to distinguish between sudden stratospheric warming (SSW) events and non-SSW events, leveraging a large ensemble of outcomes from a global climate model. Initially, I focused on a one-dimensional task, using the stratospheric zonal wind profile of SSW events encircling the 60°N latitude to train neural networks with varying complexities: logistic regression network, shallow neural network, and deep neural network. Remarkably, all the neural networks demonstrated the ability to accurately identify SSW events with a fairly high level of precision.

Next, I explore the interpretability of how these neural networks learn to differentiate between SSW and non-SSW events. To achieve this, I mask out the zonal wind fields with longitudinal windows of varying lengths to assess whether the spatial structure of disrupted winds plays a crucial role in the network's learning process. Interestingly, when the masking window is short, neither the shallow nor deep neural networks exhibit apparent spatial dependence. In contrast, the logistic regression network displays a more pronounced spatial dependence, centered around 160°W, where there is minimal variation and a negative mean value of zonal wind. As the window length increases, both the shallow and deep neural networks begin to exhibit spatial dependence in their learning. This suggests that these networks progressively take into account larger spatial patterns of the disrupted winds to make distinctions between SSW and non-SSW events.

To further investigate the two-dimensional spatial dependence, I trained a convolutional neural network (CNN) by utilizing the two-dimensional zonal wind fields across the Northern Hemisphere. Similar mask-out tests were performed, strategically masking out rectangular regions with varying sizes. The spatial dependence observed in

doi:10.6342/NTU202303879

the two-dimensional neural network closely aligns with that of the one-dimensional networks, with a notable focus on the region with negative zonal winds. However, the spatial extents expanded further north and south of the 60°N latitude. In addition to the zonal wind profile, I also incorporated two-dimensional geopotential height fields at 10 hPa to train the CNN and repeated the mask-out analysis. Two regions of high spatial dependence emerged prominently:

1. The region where large negative zonal wind fields were located, consistent with the results obtained from training using zonal wind profiles alone. This finding highlighted the geostrophic wind balance between wind and geopotential height fields.

2. A circular patch centered around the North Pole, primarily associated with the largest geopotential height difference between SSW and non-SSW events. The CNN model, therefore, not only exploited the geostrophic relationship but also leveraged the significant contrast between SSW and non-SSW events to effectively learn the categorization task.

Additionally, I investigated the effect of specifically shielding zonal wind fields at different latitudes (55°N, 60°N, 65°N, and 75°N). The strongest spatial dependence occurs around 160°W at 60°N and decreases with increasing latitude. Interestingly, when the 55°N zonal winds were masked, CNN still exhibits some misclassification rate. This indicates that in addition to the 60°N on which SSW is defined, the information around 55°N is important for identifying SSW events.

Overall, the two-dimensional CNN approach, utilizing both zonal wind fields and geopotential height fields, provided further insights into the spatial patterns associated

doi:10.6342/NTU202303879

with SSW events, shedding light on the interplay between wind and geopotential height fields in their classification.

The aforementioned results underscore the effectiveness of interpretable deep learning tools in learning the spatial information related to SSW events and uncovering their spatial dependence. These findings carry crucial implications for predicting SSW genesis and understanding the subsequent surface impacts. The ability of the deep learning models to capture and interpret the complex spatial patterns associated with SSW events opens up promising avenues for advancing our predictive capabilities in this area. By leveraging this spatial information, we can enhance our understanding of the processes leading to SSW events and their potential impacts on the Earth's surface, including weather patterns and climate variability. This, in turn, may lead to improved prediction and early warning systems for SSW events, which can have significant implications for various sectors, including weather forecasting, climate research, and societal preparedness.

doi:10.6342/NTU202303879

APPENDIX

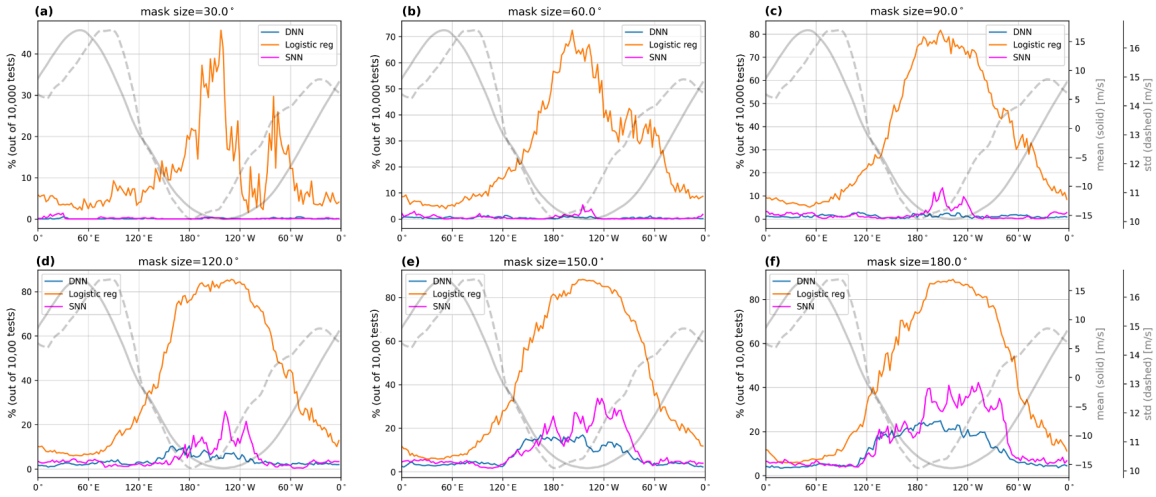


Figure 12: The error SSW classification rates of logistic regression network, SNN and DNN as a function of the central longitudes of the masks. (a)-(f) represent the masked longitudinal window of 30°, 60°, 90°, 120°, 150°, 180°, respectively. The networks are trained with EXP2 and tested with EXP1 data. The mean (solid) and standard deviation (dashed) of 60°N, 10 hPa zonal winds of 759 EXP1 displacement-type SSW events are also shown.

doi:10.6342/NTU202303879

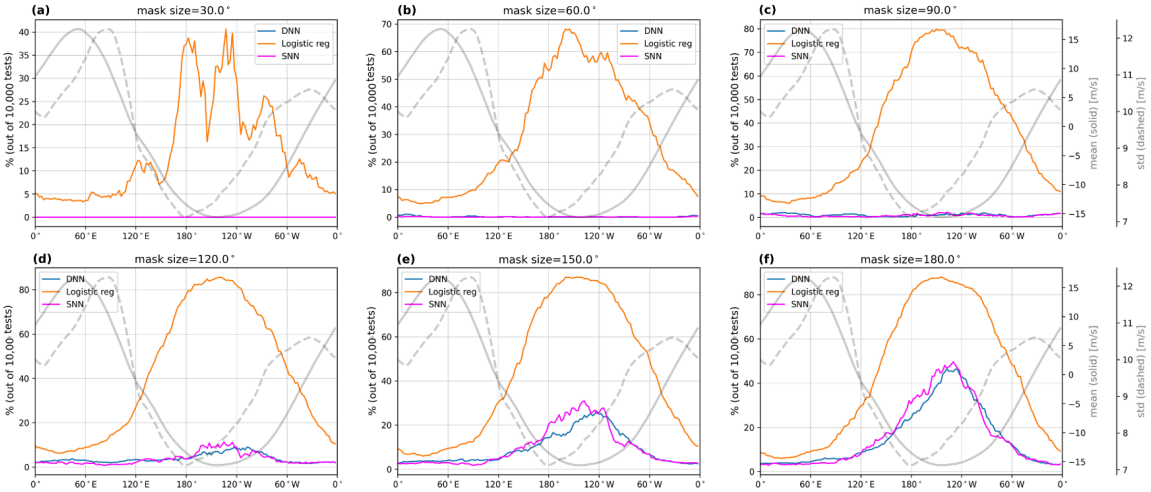


Figure 13: The error SSW classification rates of logistic regression network, SNN and DNN as a function of the central longitudes of the masks. (a)-(f) represent the masked longitudinal window of 30°, 60°, 90°, 120°, 150°, 180°, respectively. The networks are trained and tested with a mixture of EXP1 and EXP2 data. The mean (solid) and standard deviation (dashed) of 60°N, 10 hPa zonal winds of the 769 SSW events used for testing are also shown.

doi:10.6342/NTU202303879

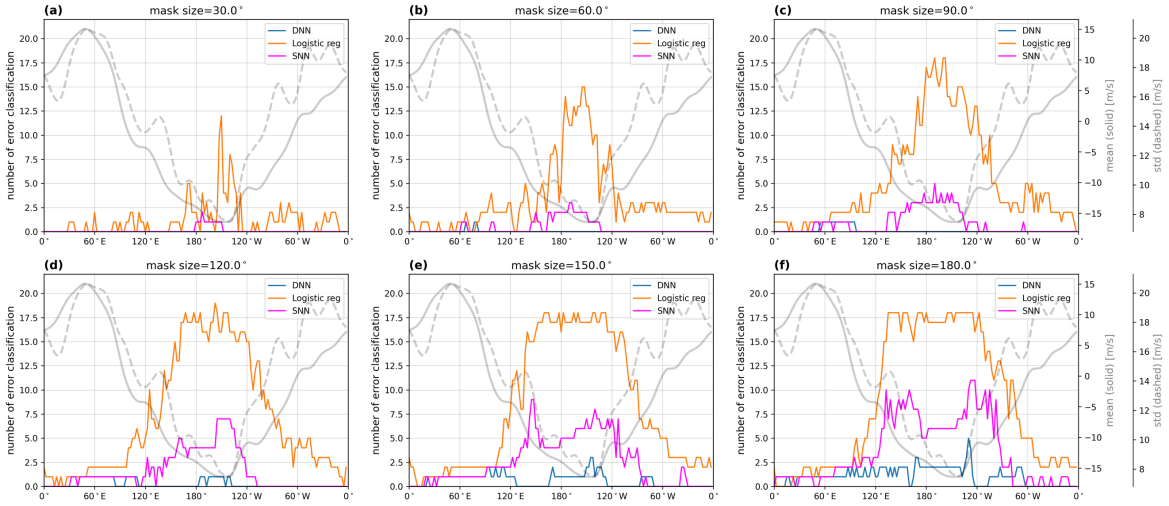


Figure 14: The misclassification counts of 22 SSW events from the NCEP/NCAR reanalysis for logistic regression network, SNN and DNN as a function of the central longitudes of the masks. (a)-(f) represent the masked longitudinal window of 30°, 60°, 90°, 120°, 150°, 180°, respectively. The networks are trained with EXP1 and tested with NCEP/NCAR reanalysis data. The mean (solid) and standard deviation (dashed) of 60°N, 10 hPa zonal winds of the 22 SSW events used for testing are also shown.

doi:10.6342/NTU202303879

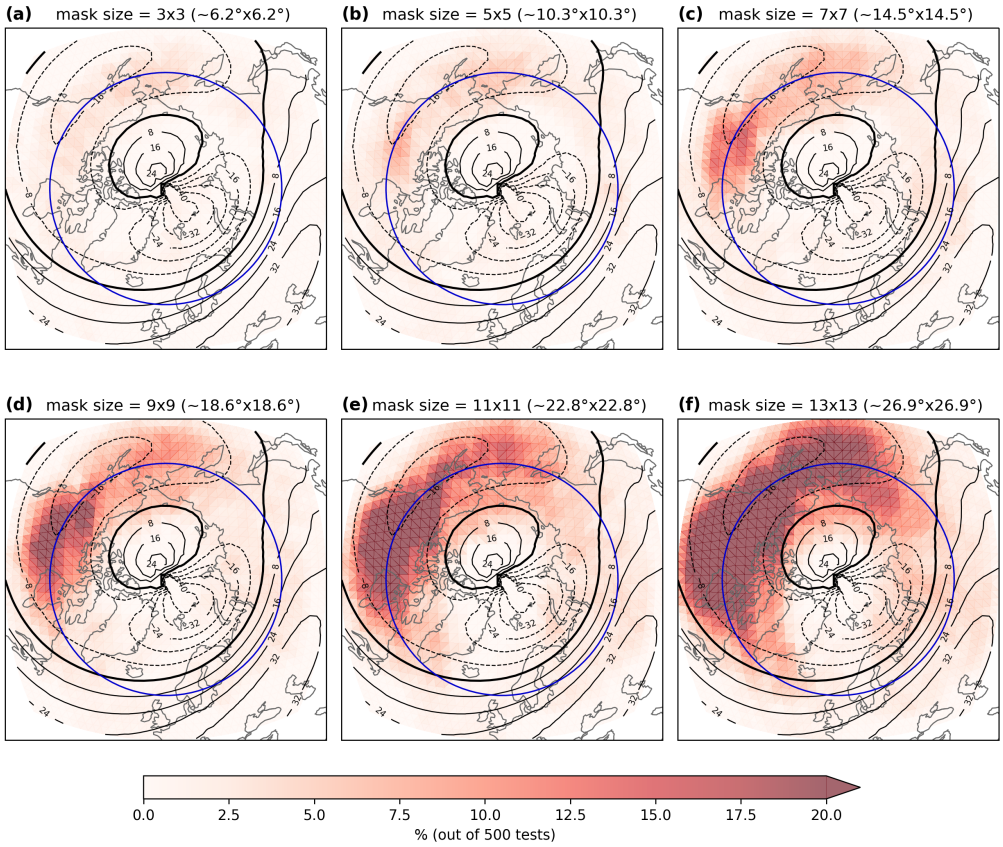


Figure 15: Red patches: the error displacement-type SSW classification rates of zonal wind fields with the central grids of the rectangular masks. (a)-(f) represent rectangular masking regions with side length of 6.2°, 10.3°, 14.5°, 18.6°, and 22.8°, respectively. Contour lines: the mean 10 hPa zonal wind field (in m/s) of 589 EXP2 displacement-type SSW events. The longitude 60°N is marked with a blue circle in each map.

doi:10.6342/NTU202303879

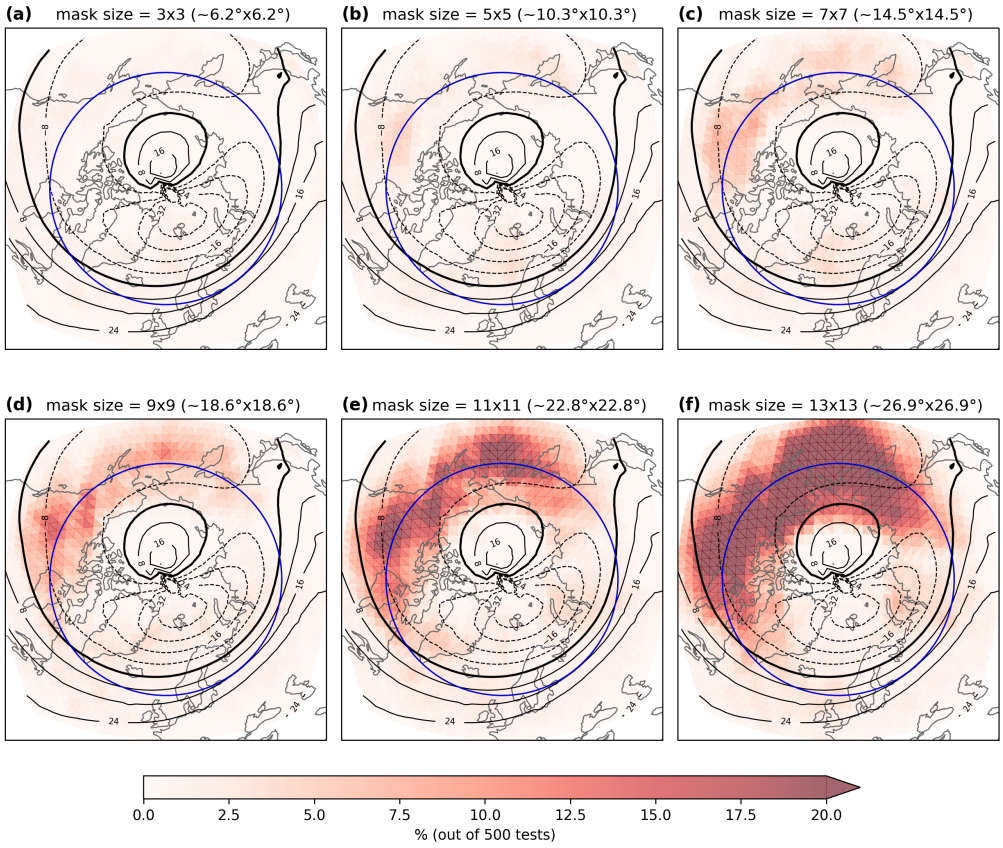


Figure 16: Red patches: the error splitting-type SSW classification rates of zonal wind fields with the central grids of the rectangular masks. (a)-(f) represent rectangular masking regions with side length of 6.2°, 10.3°, 14.5°, 18.6°, and 22.8°, respectively. Contour lines: the mean 10 hPa zonal wind field (in m/s) of 191 EXP2 splitting-type SSW events. The longitude 60°N is marked with a blue circle in each map.

doi:10.6342/NTU202303879

REFERENCES

[1] Andrews, D. G., Taylor, F. W., & McIntyre, M. E. (1987). The Influence of Atmospheric Waves on the General Circulation of the Middle Atmosphere [and Discussion]. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *323*(1575), 693–705. http://www.jstor.org/stable/38143

[2] Baldwin, M. P., Ayarzagüena, B., Birner, T., Butchart, N., Butler, A. H., Charlton-Perez, A. J., Domeisen, D. I. V., Garfinkel, C. I., Garny, H., Gerber, E. P., Hegglin, M. I., Langematz, U., & Pedatella, N. M. (2021). Sudden Stratospheric Warmings. *Reviews of Geophysics*, *59*(1), [e2020RG000708]. https://doi.org/10.1029/2020RG000708

[3] Baldwin, M. P., and Dunkerton, T. J. (1999), Propagation of the Arctic Oscillation from the stratosphere to the troposphere, *J. Geophys. Res.*, 104(D24), 30937–30946, doi:10.1029/1999JD900445.

[4] Baldwin, M. P., & Dunkerton, T. J. (2001). Stratospheric harbingers of anomalous weather regimes. *Science*, *294*(5542), 581-584. DOI:10.1126/science.1063315

[5] Baldwin, M. P., Stephenson, D. B., Thompson, D. W., Dunkerton, T. J., Charlton, A. J., & O'Neill, A. (2003). Stratospheric memory and skill of extended-range weather forecasts. *Science*, *301*(5633), 636-640. doi: 10.1126/science.1087143

[6] Barnes, E. A., Barnes, R. J., Martin, Z. K., & Rader, J. K. (2022). This Looks Like That There: Interpretable neural networks for image tasks when location matters. *Artificial Intelligence for the Earth Systems*, *1*(3), e220001. doi: https://doi.org/10.1175/AIES-D-22-0001.1

doi:10.6342/NTU202303879

[7] Barnes, E. A., Toms, B., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2020). Indicator patterns of forced change learned by an artificial neural network. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2020MS002195. doi: https://doi.org/10.1029/2020MS002195

[8] Barriopedro, D., & Calvo, N. (2014). On the Relationship between ENSO, Stratospheric Sudden Warmings, and Blocking, *Journal of Climate*, *27*(12), 4704-4720. doi: https://doi.org/10.1175/JCLI-D-13-00770.1

[9] Blume, C., Matthes, K., & Horenko, I. (2012). Supervised learning approaches to classify sudden stratospheric warming events. *Journal of the atmospheric sciences*, *69*(6), 1824-1840. doi: https://doi.org/10.1175/JAS-D-11-0194.1

[10] Butler, A. H., Seidel, D. J., Hardiman, S. C., Butchart, N., Birner, T., & Match, A. (2015). Defining Sudden Stratospheric Warmings, *Bulletin of the American Meteorological Society*, *96*(11), 1913-1928. doi: https://doi.org/10.1175/BAMS-D-13-00173.1

[11] Butler, A. H., Sjoberg, J. P., Seidel, D. J., & Rosenlof, K. H. (2017). A sudden stratospheric warming compendium. *Earth System Science Data*, *9*(1), 63-76. https://doi.org/10.5194/essd-9-63-2017, 2017.

[12] Cámara, A. d. l., Birner, T., & Albers, J. R. (2019). Are Sudden Stratospheric Warmings Preceded by Anomalous Tropospheric Wave Activity?, *Journal of Climate*, *32*(21), 7173-7189. doi: https://doi.org/10.1175/JCLI-D-19-0269.1

[13] Charlton, A. J., & Polvani, L. M. (2007). A New Look at Stratospheric Sudden Warmings. Part I: Climatology and Modeling Benchmarks, *Journal of Climate*, *20*(3), 449-469. doi: https://doi.org/10.1175/JCLI3996.1

doi:10.6342/NTU202303879

[14] Charney, J. G., and Drazin, P. G. (1961), Propagation of planetary-scale disturbances from the lower into the upper atmosphere, *J. Geophys. Res.*, 66(1), 83– 109, doi:10.1029/JZ066i001p00083.

[15] Cohen, J., & Jones, J. (2011). Tropospheric Precursors and Stratospheric Warmings, *Journal of Climate*, *24*(24), 6562-6572. doi: https://doi.org/10.1175/2011JCLI4160.1

[16] Dai, Y., & Hitchcock, P. (2021). Understanding the Basin Asymmetry in Surface Response to Sudden Stratospheric Warmings from an Ocean–Atmosphere Coupled Perspective, *Journal of Climate*, *34*(21), 8683-8698. doi: https://doi.org/10.1175/JCLI-D-21-0314.1

[17] Dai, Y., Hitchcock, P., & Simpson, I. R. (2023). Dynamics and Impacts of the North Pacific Eddy-Driven Jet Response to Sudden Stratospheric Warmings, *Journal of Climate*, *36*(3), 865-884. doi: https://doi.org/10.1175/JCLI-D-22-0300.1

[18] Davis, N. A., Richter, J. H., Glanville, A. A., Edwards, J., & LaJoie, E. (2022). Limited surface impacts of the January 2021 sudden stratospheric warming. *Nature communications*, *13*(1), 1136. doi: https://doi.org/10.1038/s41467-022-28836-1

[19] de Fondeville, R., Wu, Z., Székely, E., Obozinski, G., & Domeisen, D. I. (2022). Improved extended-range prediction of persistent stratospheric perturbations using machine learning. *Weather and Climate Dynamics Discussions*, *2022*, 1-39. doi: https://doi.org/10.5194/wcd-4-287-2023

[20] DeGrave, A. J., Janizek, J. D., & Lee, S. I. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, *3*(7), 610-619. doi: https://doi.org/10.1038/s42256-021-00338-7

doi:10.6342/NTU202303879

[21] Domeisen, D. I., & Butler, A. H. (2020). Stratospheric drivers of extreme events at the Earth's surface. *Communications Earth & Environment*, *1*(1), 59. doi: https://doi.org/10.1038/s43247-020-00060-z

[22] Garfinkel, C. I., Feldstein, S. B., Waugh, D. W., Yoo, C., & Lee, S. (2012). Observed connection between stratospheric sudden warmings and the Madden-Julian Oscillation. *Geophysical Research Letters*, *39*(18). doi: https://doi.org/10.1029/2012GL053144

[23] Gettelman, A., Mills, M. J., Kinnison, D. E., Garcia, R. R., Smith, A. K., Marsh, D. R., ... & Randel, W. J. (2019). The whole atmosphere community climate model version 6 (WACCM6). *Journal of Geophysical Research: Atmospheres*, *124*(23), 12380-12403. doi: https://doi.org/10.1029/2019JD030943

[24] Gopinath, D. (2021). The Shapley Value for ML Models. *Towards Data Science*.

[25] Hall, R. J., Mitchell, D. M., Seviour, W. J. M., & Wright, C. J. (2021). Tracking the stratosphere-to-surface impact of Sudden Stratospheric Warmings. *Journal of Geophysical Research: Atmospheres*, 126, e2020JD033881. https://doi.org/10.1029/2020JD033881

[26] Hannachi, A., Mitchell, D., Gray, L., & Charlton-Perez, A. (2011). On the use of geometric moments to examine the continuum of sudden stratospheric warmings. *Journal of the Atmospheric Sciences*, *68*(3), 657-674. doi: https://doi.org/10.1175/2010JAS3585.1

[27] Hitchcock, P., and Haynes, P. (2016), Stratospheric control of planetary waves, *Geophys. Res. Lett.*, 43, 11,884– 11,892, doi:10.1002/2016GL071372.

[28] Hitchcock, P., Shepherd, T. G., & Manney, G. L. (2013). Statistical characterization of Arctic polar-night jet oscillation events. *Journal of Climate*, *26*(6), 2096-2116. doi: https://doi.org/10.1175/JCLI-D-12-00202.1

doi:10.6342/NTU202303879

[29] Hitchcock, P., & Simpson, I. R. (2014). The Downward Influence of Stratospheric Sudden Warmings, *Journal of the Atmospheric Sciences*, *71*(10), 3856-3876. doi: https://doi.org/10.1175/JAS-D-14-0012.1

[30] Jung, J.-H., Konor, C. S., & Randall, D. (2019). Implementation of the vector vorticity dynamical core on cubed sphere for use in the Quasi-3-D Multiscale Modeling Framework. *Journal of Advances in Modeling Earth Systems*, 11, 560–577. https://doi.org/10.1029/2018MS001517

[31] Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., ... & Joseph, D. (2018). The NCEP/NCAR 40-year reanalysis project. In *Renewable Energy* (pp. Vol1_146-Vol1_194). Routledge.

[32] Kennedy, John; Titchner, Holly; Rayner, Nick; Roberts, Malcolm. **(2017)**. *input4MIPs.MOHC.SSTsAndSeaIce.HighResMIP.MOHC-HadISST-2-2-0-0-0.* doi:10.22033/ESGF/input4MIPs.1221

[33] Kidston, J., Scaife, A. A., Hardiman, S. C., Mitchell, D. M., Butchart, N., Baldwin, M. P., & Gray, L. J. (2015). Stratospheric influence on tropospheric jet streams, storm tracks and surface weather. *Nature Geoscience*, *8*(6), 433-440. doi: https://doi.org/10.1038/ngeo2424

[34] Kolstad, E.W., Breiteig, T. and Scaife, A.A. (2010), The association between stratospheric weak polar vortex events and cold air outbreaks in the Northern Hemisphere. Q.J.R. Meteorol. Soc., 136: 886-893. https://doi.org/10.1002/qj.620

[35] Kretschmer, M., Runge, J., & Coumou, D. (2017). Early prediction of extreme stratospheric polar vortex states based on causal precursors. *Geophysical research letters*, *44*(16), 8592-8600. doi: https://doi.org/10.1002/2017GL074696

[36] Lawrence, Z. D., & Manney, G. L. (2018). Characterizing stratospheric polar vortex variability with computer vision techniques. *Journal of Geophysical*

doi:10.6342/NTU202303879

Research: Atmospheres, *123*(3), 1510-1535. doi: https://doi.org/10.1002/2017JD027556

[37] Lehtonen, I., and Karpechko, A. Y. (2016), Observed and modeled tropospheric cold anomalies associated with sudden stratospheric warmings, *J. Geophys. Res. Atmos.*, 121, 1591– 1610, doi:10.1002/2015JD023860.

[38] Li, X., Feng, M., Ran, Y., Su, Y., Liu, F., Huang, C., ... & Guo, H. (2023). Big Data in Earth system science and progress towards a digital twin. *Nature Reviews Earth & Environment*, 1-14. doi: https://doi.org/10.1038/s43017-023-00409-w

[39] Limpasuvan, V., Thompson, D. W. J., & Hartmann, D. L. (2004). The Life Cycle of the Northern Hemisphere Sudden Stratospheric Warmings, *Journal of Climate*, *17*(13), 2584-2596. doi: https://doi.org/10.1175/1520-0442(2004)017<2584:TLCOTN>2.0.CO;2

[40] Lu, C., & Ding, Y. (2015). Analysis of isentropic potential vorticities for the relationship between stratospheric anomalies and the cooling process in China. *Science Bulletin*, *60*(7), 726-738. doi: https://doi.org/10.1007/s11434-015-0757-4

[41] Lu, C., Zhou, B., & Ding, Y. (2016). Decadal variation of the Northern Hemisphere annular mode and its influence on the East Asian trough. *Journal of Meteorological Research*, *30*(4), 584-597. doi: https://doi.org/10.1007/s13351-016-5105-3

[42] Martius, O., Polvani, L. M., and Davies, H. C. (2009), Blocking precursors to stratospheric sudden warming events, *Geophys. Res. Lett.*, 36, L14806, doi:10.1029/2009GL038776.

[43] Matsuno, T. (1971). A Dynamical Model of the Stratospheric Sudden Warming, *Journal of Atmospheric Sciences*, *28*(8), 1479-1494. doi: https://doi.org/10.1175/1520-0469(1971)028<1479:ADMOTS>2.0.CO;2

doi:10.6342/NTU202303879

[44] Mitchell, D. M., Charlton-Perez, A. J., & Gray, L. J. (2011). Characterizing the variability and extremes of the stratospheric polar vortices using 2D moment analysis. *Journal of the Atmospheric Sciences*, *68*(6), 1194-1213. doi: https://doi.org/10.1175/2010JAS3555.1

[45] Nishii, K., Nakamura, H., & Orsolini, Y. J. (2011). Geographical Dependence Observed in Blocking High Influence on the Stratospheric Variability through Enhancement and Suppression of Upward Planetary-Wave Propagation, *Journal of Climate*, *24*(24), 6408-6423. doi: https://doi.org/10.1175/JCLI-D-10-05021.1

[46] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. & Lerer, A. (2017). Automatic Differentiation in PyTorch. *NIPS 2017 Workshop on Autodiff*, .

[47] Polvani, L. M., & Waugh, D. W. (2004). Upward Wave Activity Flux as a Precursor to Extreme Stratospheric Events and Subsequent Anomalous Surface Weather Regimes, *Journal of Climate*, *17*(18), 3548-3554. doi: https://doi.org/10.1175/1520-0442(2004)017<3548:UWAFAA>2.0.CO;2

[48] Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., and Kaplan, A. (2003), Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, *J. Geophys. Res.*, 108, 4407, doi:10.1029/2002JD002670, D14.

[49] Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat, F. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195-204. doi: https://doi.org/10.1038/s41586-019-0912-1

doi:10.6342/NTU202303879

[50] Ren, R., & Cai, M. (2006). Polar vortex oscillation viewed in an isentropic potential vorticity coordinate. *Advances in Atmospheric Sciences*, *23*, 884-900. doi: https://doi.org/10.1007/s00376-006-0884-6

[51] Savage, N. (2022). Breaking into the black box of artificial intelligence. *Nature*.

[52] Schoeberl, M. R. (1978), Stratospheric warmings: Observations and theory, *Rev. Geophys.*, 16(4), 521– 538, doi:10.1029/RG016i004p00521.

[53] Scott, R. K., and Polvani, L. M. (2004), Stratospheric control of upward wave flux near the tropopause, *Geophys. Res. Lett.*, 31, L02115, doi:10.1029/2003GL017965.

[54] Scott, R. K., & Polvani, L. M. (2006). Internal Variability of the Winter Stratosphere. Part I: Time-Independent Forcing, *Journal of the Atmospheric Sciences*, *63*(11), 2758-2776. doi: https://doi.org/10.1175/JAS3797.1

[55] Shapley, L. (1953). A Value for n-Person Games. In: Kuhn, H. and Tucker, A., Eds., Contributions to the Theory of Games II, Princeton University Press, Princeton, 307-317. doi: https://doi.org/10.1515/9781400881970-018

[56] Sigmond, M., Scinocca, J. F., Kharin, V. V., & Shepherd, T. G. (2013). Enhanced seasonal forecast skill following stratospheric sudden warmings. *Nature Geoscience*, *6*(2), 98-102. https://doi.org/10.1038/ngeo1698

[57] Titchner, H. A., and Rayner, N. A. (2014), The Met Office Hadley Centre sea ice and sea surface temperature data set, version 2: 1. Sea ice concentrations, *J. Geophys. Res. Atmos.*, 119, 2864– 2889, doi:10.1002/2013JD020316.

[58] Toms, B. A., Barnes, E. A., & 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), e2019MS002002. doi: https://doi.org/10.1029/2019MS002002

doi:10.6342/NTU202303879

[59] Wu, Z., Beucler, T., Székely, E., Ball, W., & Domeisen, D. (2022). Modeling stratospheric polar vortex variation and identifying vortex extremes using explainable machine learning. *Environmental Data Science, 1*, E17. doi: https://doi.org/10.1017/eds.2022.19

doi:10.6342/NTU202303879