Clustering of polar vortex states using convolutional autoencoders

Mikhail A. Krinitskiy, Yulia A. Zyulyaeva, and Sergey K. Gulev

Shirshov Institute of Oceanology, Russian Academy of Sciences
36 Nahimovskiy pr., Moscow, 117997, Russia
krinitsky@sail.msk.ru

Abstract

A profound understanding of the stratospheric wintertime dynamics and its climate changes are important for improving seasonal forecast skill. The primary goal of the research of the wintertime Arctic stratospheric polar vortex (PV hereafter) is defining its states and their clustering. Manual classification is a highly time-consuming task suffering of researcher subjectivity. We apply deep learning methods that let us cluster the PV states based on their spatial structure. We designed the particular kind of neural networks called variational convolutional autoencoder with the sparsity constraint (SpCVAE). We applied the hierarchical agglomerative clustering algorithm to the states pf PV described by their embedded representation generated by SpCVAE. 96-dimensional embedded representation was found to be optimal with high samples reconstruction quality. The best number of clusters was chosen based on "elbow rule" and topic-specific reasoning. The approach applied let us automatically distinguish weak PVs of "displacement" and "split" types, as well as to isolate several strong vortex states of different shift directions. These results are only obtainable when one considers the spatial structure of the PV. We have constructed the calendar of the PV states based on the clustering result. Clustered events of weak PVs were examined and demonstrated good correspondence with the calendar of sudden stratospheric warmings that have been built manually. This result is now the basis for the research of the stratosphere-troposphere interaction for existing and future climate scenarios.

1 Introduction

Nowadays, skillful numerical weather prediction is limited by about 10 days due to the chaotic nature of atmospheric dynamics [1]. Skillful seasonal forecasting typically relies on the predictability of slow-varying components of the climate system, such as sea surface temperature, sea ice, snow cover, and soil moisture. For instance, the predictability of El Niño Southern Oscillation (ENSO) phenomenon is a remarkable example of high skills of the seasonal forecast system [2,3]. However, recent studies demonstrated that the maximum seasonal forecast skills have not yet been achieved, and pointed to the stratosphere as a potential source for enhancing seasonal predictability [1,4].

Before [5] the stratosphere was considered playing a passive role in the stratosphere-troposphere coupling, that is, it does not influence troposphere dynamics. Baldwin and Dunkerton showed [5] that the strength of the polar vortex affects the main features of the paths of the cyclones propagation. From that time, interest in the stratospheric dynamics and its climatic changes have constantly been growing.

Polar stratosphere became extremely cold during the polar night, and meridional temperature gradient becomes strong, which leads to the formation of the polar vortex. Planetary waves propagating from the troposphere to the stratosphere disturb and sometimes destroy polar vortex. These events are known as Sudden Stratospheric Warming events (SSW), as the temperature near the pole dramatically increase (up to 40° per 4 - 7 days) when the vortex is destroyed. Two types of SSW are classified nowadays: “displacement” with the center of the vortex shifted significantly towards the equator, and “split” with the vortex split into two vortices. There are periods of the extremely strong vortex as well. Variability of the PV intensity is the most influential factor of intraseasonal variability in the winter stratosphere. During springtime as the polar day is coming to high latitudes, temperature gradient decrease, so the polar vortex disappears. Summertime variability of the stratosphere dynamics is low, and there is no source of the stratosphere-troposphere coupling.

It has been shown that weakenings of the polar vortex precede the shift of the storm tracks (main path of mid-latitude cyclones propagation) to the south, which may cause cold outbreaks in the North Atlantic - Europe region [6–10]. These anomalies may act in the troposphere up to 2 months [6,11]. Amplitudes of these anomalies are comparable to the effect

of the ENSO [3]. Therefore, one can extend and improve long-term weather forecast [12–16]. However, relatively little attention is paid to strong polar vortex events compared to SSW. One of the main reasons for that is the difficulty of their identification classification.

There are a few metrics for the describing states of the PV [3,17]. Widely used key features of PV for applying different types of analysis are the aggregated and diagnostic parameters like maximum pressure anomaly, zonal-mean zonal winds, etc. These parameters do not preserve spatial characteristics like PV center shift from the North Pole, vortex shape parameters or various anomalies duration.

Over last decades machine learning methods demonstrated spectacular results in research of climatic changes of atmospheric circulation [18–22]. Researchers mostly rely on state-of-the-art clustering methods. Using the resulting atmosphere states grouping one can assess atmospheric circulation characteristics and trends within each cluster. Most of this kind of works are focused on troposphere states research [18–20,22]. However, machine learning techniques are shown recently to be fruitful for stratosphere states clustering [21]. Mostly Kohonen self-organizing maps [23] is applied in these works for clustering. However, its ability to preserve the topology of a dataset is rarely used. Hierarchical agglomerative clustering is the less frequently used method [18,21]. For this type of clustering, the source data of geophysical fields are usually aggregated following a researcher sense of a particular operation ability to preserve the informational content. As a result, each PV state is represented with vectors. This procedure leads to the loss of crucial information about the spatial configuration of objects being researched.

In our work, we focused on polar vortex states research with the use of geopotential height at 10hPa level (see section 2.1 “Data and preprocessing”). We applied hierarchical agglomerative clustering method on the embedded representations of PV states that are generated by variational convolutional autoencoder with the sparsity constraint (hereafter SpCVAE) [24,25]. SpCVAE as a feature extractor and a tool for dimensionality reduction preserves features of PV spatial configuration, yet it is capable of decreasing the computational costs of clustering. Purposes of a SpCVAE here are dimensionality reduction and extraction of significant information based on the whole PV states dataset. Being a neural network trained end-to-end the SpCVAE avoids manual feature engineering. Thus there is no need to rely on a researcher`s sense of the importance of PV aggregated parameters.

2 Data and methods

2.1 Data and preprocessing

We analyzed geopotential heights (HGT) and potential vorticity (PVt) fields at 10 hPa level from JRA-55 (Japanese 55-year Reanalysis) in this study [26]. We considered wintertime period December-February (DJF) for 1958 - 2014 years. The spatial resolution of the data is 1,25° x 1,25°; the upper level of the model is 0.1 hPa, which is crucial for the analysis of stratospheric processes. JRA-55 shown to be in good coherence with all modern reanalysis data (S-RIP [27]). The main advantage of JRA-55 is the extended period from the 1958 year in comparison with ERA-Interim (European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis data) and MERRA (NASA Modern Era Reanalysis for Research and Applications) that start from 1979.

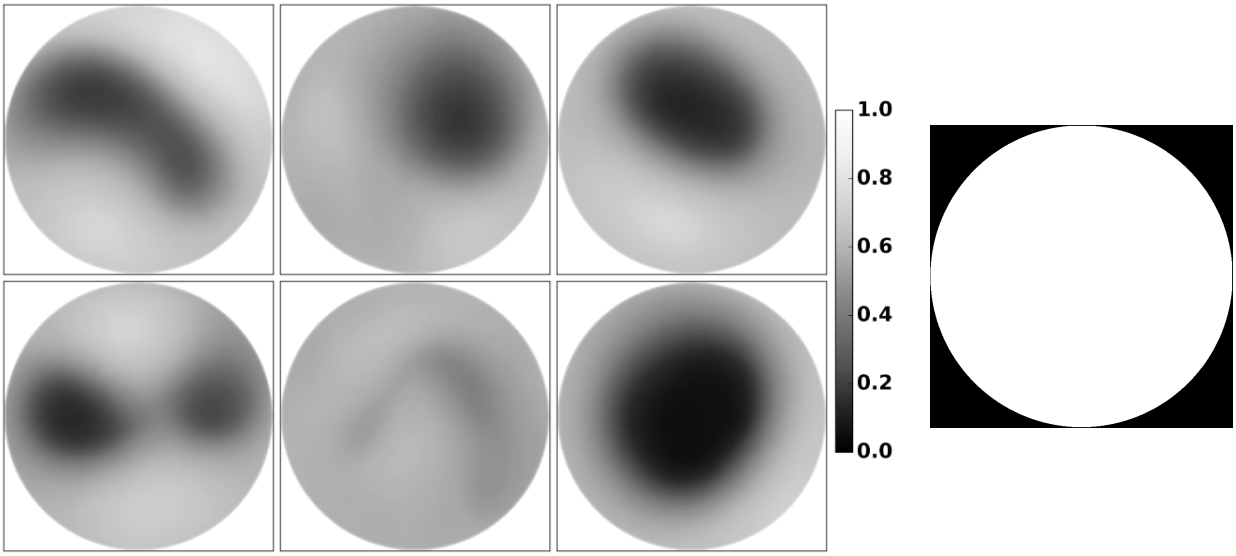
Source data for describing PV states were considered to be field values to the north from 40°N.

At the preprocessing stage, timestep-defined snapshots of HGT and PVt fields were projected using North-polar Lambert azimuthal projection and interpolated to form a two-dimensional flat matrix of size 256x256. For each date of a year, we calculated the climatological median and subtracted it from each snapshot of this date (e.g., from all snapshots within January 25th of each year). Since we use the North-polar Lambert azimuthal projection, only central rounded part of each sample is informative, so during all calculations, we used the mask M_{ij} (fig. 1b).

We further normalized these snapshots, so all their values are limited and take values between 0.0 and 1.0:

$$x_{ij} = \frac{x_{ij} - \min\limits_{i,j} X}{\max\limits_{i,j} X - \min\limits_{i,j} X}, \tag{1}$$

where X denotes the whole dataset snapshots; x is a particular snapshot, i and j are x matrix indices; $\min(X)$ and $\max(X)$ are calculated taking the mask M_{ij} into account. This kind of normalization applied to both HGT and PVt datasets separately. With this preprocessing procedure, the PV states dataset is represented with two fields containing 21476 matrices of size 256 x 256 which values fit the range between 0.0 and 1.0. Some examples of this dataset are shown in fig. 1a.



(a) (b)

Figure 1: (a) Examples from the dataset of PV states (HGT values only, normalized); (b) mask M_{ij}

2.2 Sparse variational convolutional autoencoder

Convolutional autoencoder (CAE) is a particular type of neural autoencoders [24,25] which purpose is to deal with two-dimensional fields when it is crucial to take into account spatial relations between data anomalies. The fundamental purpose of autoencoders is the approximate copying of input data with some constraints imposed on the network structure. These constraints are often the restriction on the maximal dimensionality of the embedded representation obtained on the encoder output (see fig. 2), and the CAE is called undercomplete. In this sense, the whole undercomplete CAE is essentially a transformation: $\mathcal{F}: A \to H \to \tilde{A}$, where A is an input example matrix, $A \in \mathbb{A}$; H is its hidden representation vector such as $H \in \mathcal{H}$ where $\mathcal{H}$ is a hidden representations space, and $\tilde{A}$ is the reconstructed example matrix, $\tilde{A} \in \mathbb{A}$. Here $\mathbb{A}$ is the space $\mathbb{R}^n$ where n is the number of pixels of input examples. In our study, $n = 2 * 256 * 256$ since HGT and PVt projected examples are matrices 256x256. The transformations $A \to H$ and $H \to \tilde{A}$ are referred hereafter as encoder and decoder respectively. An autoencoder is trained with the loss function defined according to the similarity definition suitable to the problem. Mean squared error is a widely used loss function that is suitable for most tasks that imply the processing of geophysical fields. In a case of limited source data values, one may normalize them accordingly to use binary cross-entropy loss (BCE) (eq. 2). We normalized source data, so its values are limited (see Section 2.1) therefore we use BCE loss $\mathcal{L}_{bce}$:

$$\mathcal{L}_{bce}(X, p) = -\sum_{i=0}^{m}\sum_{j=0}^{n} M_{ij}\left(x_{ij} \ln \mathcal{F}\left(x_{ij}\right)\right) , \tag{2}$$

It was shown [24,25,28] that successful training of a neural network implies tuning its weights the way that the trainable part of the network extracts latent parameters distribution of the training dataset. With the trained CAE, the embedded representation [29,30] of input samples preserve enough information for the network to be able to reconstruct it with the appropriate quality. We use this feature of CAEs to perform dimensionality reduction with minimum loss of information about latent parameters variability and the spatial structure of PV states.

We applied commonly used techniques of convolutional neural networks quality improvement called Transfer Learning (hereafter TL) [28,31–35] and Fine Tuning (hereafter FT) [36,37]. In practice, TL implies the construction of a new neural network based on a subset of layers of the network that was previously trained on a huge dataset, e.g., ImageNet [38]. Moreover, these layers are set to be "frozen," that is, their weights are not optimized during training. This approach is fruitful when one uses a dataset which statistical characteristics are similar to ImageNet. In practice, it means that the new dataset images should contain visual patterns similar to ones that are frequently met in ImageNet. TL application significantly decreases the computational costs of new models training. Fine Tuning approach implies turning off the "frozen" state for some top layers of the transferred neural network. With this approach, one can tune the whole CAE taking into account the peculiar properties of the dataset.

We applied the TL technique while building the encoder part of the CAE. We used pre-trained VGG-16 [39] as a transferred network. We used the convolutional part of VGG-16 and a set of new fully-connected layers attached to it. VGG-16 convolutional sub-network is denoted as "convolutional core" in fig. 2. We also applied several regularizations

to prevent overfitting. Particularly we used the dropout [40] approach and L2 regularization which penalizes high-value weights.

In our model, the VGG-16 output is reshaped to a vector which is then the input for the encoder fully-connected part. This fully-connected part consists of three layers FC1, FC2, and FC3, which are alternating with dropout layers. FC3 is an intermediate one between encoder and decoder. The output of FC3 is the hidden representation of the input sample and the input vector for the decoder.

Autoencoders usually tend to be symmetric. With this approach, we constructed the decoding part to be mirrored to the encoder. All the weights of the decoder are trainable. Following the best practices of composing convolutional autoencoders, we used two-dimensional upsampling layers which mirror max-pooling layers of the encoder. Upsampling here is an operation of repeating the layer`s input rows and columns by the specified number of times. Decoder outputs are the reconstructed HGT and PVt fields of input example which has to be similar to the input in a sense defined by the loss function of the network which is BCE loss (eq. 2.)

The dimensionality constraint mentioned above is the dimensionality of $\mathcal{H}$. The common approach of the clustering involving autoencoders relies on their capability of projecting the input examples to the hidden representation space $\mathcal{H}$. This transformation was shown to be trained so that the examples which are close to each other in $\mathcal{H}$ are similar [29]. However, the key feature of a dataset should be the opposite for the reliable, stable, and reproducible clustering, that is, similar examples should be located close to each other in $\mathcal{H}$. The ordinary undercomplete CAE does not guarantee this property of the projection $\mathbb{A} \to \mathcal{H}$. This issue might be addressed with the variational autoencoder (VAE) [41] which was shown to produce continuous latent variables space $\mathcal{H}$. That is, with VAE involved, similar examples are located close to each other in $\mathcal{H}$. However, since the distribution of the features of H is normal in case of VAE, the clustering cannot produce valuable results in the generated feature space $\mathcal{H}$. In our study, this issue is addressed with the constraint of sparsity, that is, features of the vector H are forced to be Bernoulli-distributed. With this constraint, the hidden representation vectors H tend to be sparse, that is, only a few features are non-zero for a particular example A. The undercomplete CAE with the mentioned constraints imposed is referred hereafter as sparse variational convolutional autoencoder (SpCVAE). Its structure is presented in fig. 2. As shown in fig. 2, the fields of the input example are processed separately, similar to the approach applied in [42].

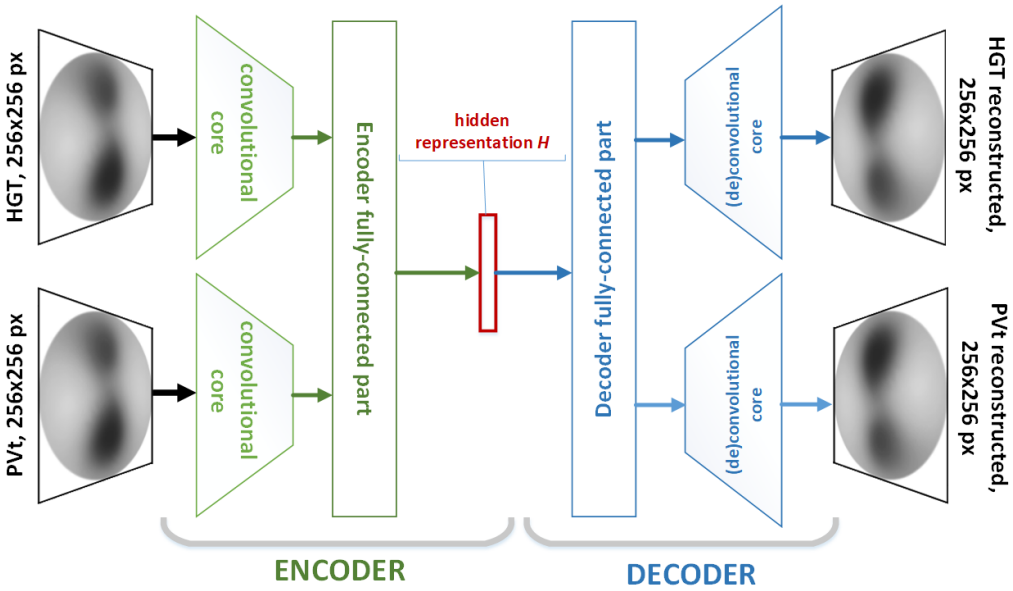


Figure 2: The structure of the sparse variational convolutional autoencoder.

In our work, the only hyperparameter of the proposed SpCVAE is the number of nodes of the FC3 layer. This number at the same time is the number of features of the hidden representation H (see fig. 2) and the dimensionality of $\mathcal{H}$ (hereafter $HDim$). There is a trade-off between reconstruction quality and the $HDim$. We use the multiscale structural similarity ($MSSSIM$ [43]) as a measure for reconstruction quality. For optimization reasons, we use the metric ($1 - MSSSIM$) with the rule "less is better" (fig. 3a). We have conducted the research of the reconstruction quality versus the $HDim$ (see fig. 3a). There is a reasonable value of $HDim = 96$ since after this value the reconstruction quality stops improving significantly. There are more candidates for the best choice of $HDim$, however, only starting with the $HDim = 96$ the clustering results become stable and reproducible. Taking this result into account, we further used the SpCVAE with 96 neurons of layer FC3. We have trained this SpCVAE using the data described in section 2.1. We use then the encoder output in the inference mode of the trained SpCVAE as PV states representation of reduced dimensionality.

2.3 Hierarchical agglomerative clustering

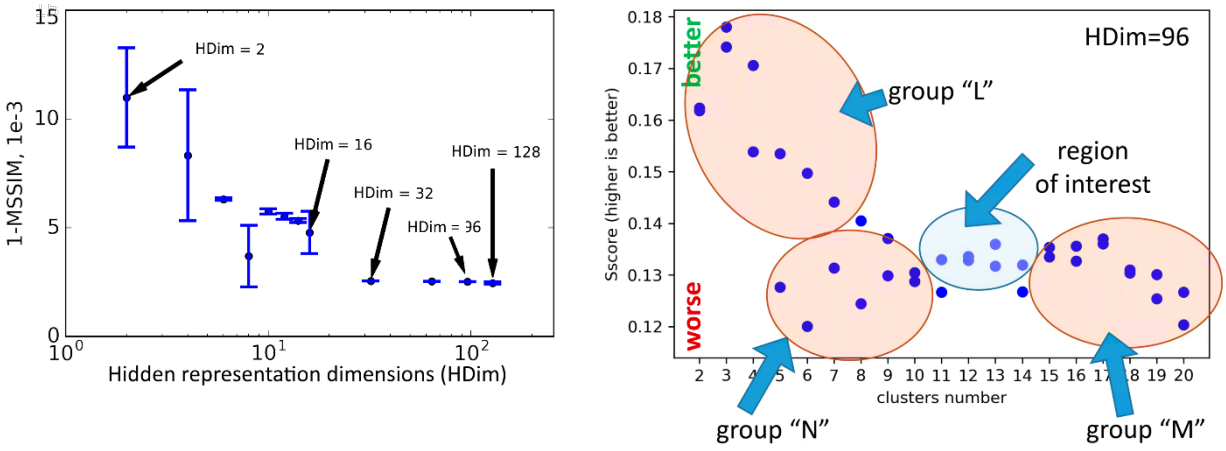
We applied Lance-Williams hierarchical agglomerative clustering [44–46] to define groups of stable PV states. This method is frequently used for atmosphere and stratosphere states clustering [18–22]. We applied this clustering method to PV states objects described with low-dimensional hidden representations generated by SpCVAE (see fig. 2). We considered the Euclidean metric as a distance between objects in this feature space. We used Ward minimal inter-cluster distance [47] as a criterion for clusters union. Ward inter-cluster distance between clusters U and V is the following:

$$D^W(U,V) = \frac{\|U\|\|V\|}{\|U\| + \|V\|}\rho^2\left(\sum_U \frac{x_U}{\|U\|}, \sum_V \frac{x_V}{\|V\|}\right), \tag{3}$$

where x_U and x_V are embedded representation vectors for objects assigned to clusters U and V respectively; ρ denotes Euclidean distance between vectors; $\|U\|$ and $\|V\|$ are elements number of clusters U and V. Hierarchical agglomerative clustering algorithm for a set of objects $x_i, i = 1 \dots n$ is represented with pseudocode:

1. $step = 1$; initialize the starting set S_1 – the universal set of one-element clusters $\{\{x_1\},\{x_2\},\dots,\{x_n\}\}$; Ward inter-cluster distances calculated using eq. 3 are equal to halved element-to-element squared Euclidean distances;
2. for each $step = 2 \dots n$ repeat:
   a. search for a pair of most close clusters in terms of Ward inter-cluster distances (eq. 3):
   $$(U,V)_t = \underset{U,V \in S_t}{\operatorname{argmin}} D^W(U,V) \ . \tag{4}$$
   b. unite U and V, exclude U and V from S_t, add the united cluster to S_t:
   $$W = U \cup V, \tag{5}$$
   $$S_{t+1} = (S_t \backslash \{U,V\}) \cup W. \tag{6}$$
   c. for each $C \in S_{t+1}$, calculate inter-cluster distances to $D^W(C,W)$ using eq. 3.

Agglomerative hierarchical clustering procedure with Ward inter-cluster distance definition (eq. 3) is the one demonstrated most valuable results in a set of synthetic clustering problems [48]. With this procedure, the only hyperparameter is the target clusters number K. We use the empirical "elbow rule" to choose the best clusters number. Additionally, we considered topic-specific reasoning: we wanted the clustering method to be capable of discriminating weak PV states of types "displacement" and "split" yet to be capable of discriminating various shifted states that were demonstrated recently to be present [21].



(a) (b)

Figure 3: (a) $(1 - MSSSIM)$ as a measure of reconstruction quality versus $HDim$ (lower is better); (b) mean silhouette score (higher is better) versus clusters number with the fixed encoder which is producing lower-dimensional representations H of $HDim = 96$.

We considered mean Silhouette score (hereafter $Sscore$) as a measure for clustering quality and used it for "elbow rule". We calculated $Sscore$ using Euclidean metric of the hidden representation feature space $\mathcal{H}$. For each i-th object of the j-th cluster $x_i \in C_j$: $a(x_i)$ is the mean distance between x_i and all other objects of cluster C_j; $b(x_i)$ is the mean distance between x_i and all the objects of all other clusters:

$$a(x_i) = \frac{1}{\|C_j\|}\sum_{x_k \in C_j, k \neq i} D(x_i, x_k) \ , \tag{7}$$

$$b(x_i) = \frac{1}{\|\mathcal{T}\| - \|C_j\|} \sum_{x_k \in \mathcal{T} \setminus C_j} D(x_i, x_k) \ , \tag{8}$$

where $\mathcal{T}$ is the whole dataset, $\|\mathcal{T}\|$ is the number of its elements, C_j is the cluster that x_i is assigned to, $\|C_j\|$ is the number of its elements; $D(x_i, x_k)$ is the function defining the distance between x_i and x_k. $D(x_i, x_k)$ is the Euclidean distance in our study. With these notations $Sscore$ for one object x_i is given by:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\big(b(x_i), a(x_i)\big)} \ , \tag{9}$$

and mean $Sscore$ is given by:

$$S(\mathcal{T}) = \frac{1}{\|\mathcal{T}\|} \sum_{\mathcal{T}} s(x_i) \ . \tag{10}$$

$Sscore$ for each x_i is the measure of its similarity to the cluster C_j and yet its dissimilarity to all the other elements outside C_j. So the higher $Sscore$, the more x_i is similar to C_j and the less similar to other clusters. Therefore, mean Silhouette score (eq. 10) is considered as a measure of clustering quality with the rule "higher is better". Even though the maximum mean $Sscore$ is achieved with two clusters, more reasoning should be involved when one is choosing the number of clusters. First, there should be observed "split" and "displacement" SSW events. There are also should be observed at least one strong pole-centered state, and some other shifted states which were presented recently [21]. For each number of clusters more than 3 we inspected maps of mean geopotential heights at 10hPa level. We have chosen the minimal number of clusters that let us discriminate weak states of PV of types "displacement" and "split". This discrimination is observed starting from $K = 7$. With this reasoning, the group "L" of clustering results (see fig. 3b) should not be considered as an option. There is also group "N" of clustering results which are characterized by low mean Sscore, that is, low clustering quality. The results within the "M" group are characterized by too high clusters number or low clustering quality. In our study, the group "region of interest" is considered as a group of promising clustering results (fig. 3b). The numbers of clusters which produces high average $Sscore$ within this group are 12 and 13. In our study, we use $K = 12$.

Summarizing the proposed method for clustering states of PV here is its general structure:

1. Prepare and preprocess PV states data (HGT and PVt fields);
2. Construct sparse variational convolutional autoencoder (fig. 2); train this SpCVAE on prepared PV states dataset;
3. Apply dimensionality reduction using trained SpCVAE. Representation of the reduced dimensionality is the encoder output for each PV state presented to SpCVAE as input example;
4. Apply hierarchical agglomerative clustering;
5. Choose the best hidden representation dimensionality $HDim$ and best clusters number K based on the metrics of examples reconstruction $MSSSIM$, clustering quality $Sscore$, stability and reproducibility of clustering, and additional problem-specific reasoning (fig. 3).

3 Results and discussion

We applied the approach presented in Section 2 to the dataset described in Section 2.1, "Data and preprocessing." PV states were clustered using $K = 12$ number of clusters. We calculated the map of mean geopotential heights at 10hPa level for each cluster. These maps are shown in fig. 4. We also selectively inspected individual PV states represented by HGT fields. Visual inspection of these examples shows that PV states grouped by the proposed method are similar to each other within each cluster. SSW events are clearly seen in the composites in fig. 4: cluster 2 for "split" type and clusters 1 and 3 for "displacement" type.

We have conducted a more detailed analysis of the clusters 1-3. The calendar in fig. 6 presents the periods associated with the states of the clusters 1-3 and concurrently the known SSW events identified by experts. Central dates of the "split" and "displacement" SSWs obtained by Charlton and Polvani for period 1958 – 2002 [49], and non-classified SSWs from [50] for period 2003 - 2013 are shown. Almost all expert-defined SSW events are clearly colocated in time with the segments of clusters 1-3. Most of "split" SSW events are co-located with the cluster 2 or with the segment consisting of states associated with two clusters including cluster 2. The only one SSW event in 2002 missing corresponding states of clusters 1-3 is a subject for further research.

Central dates of the SSWs are defined as the dates when zonal-mean zonal winds at 10hPa and 60°N fall below zero m/s (became easterly). In fig. 5b, we present zonal wind averaged along 60°N for each cluster. In this figure, clusters 1-3 are distinctively characterized by winds close to zero or even easterly winds. This behavior is consistent with the nature and definition of SSW events [17]. Since there was no expert-level knowledge involved during clustering, we may consider the proposed method to be capable of objective weak vortex clustering.

It is also clear that clusters 10 and 12 represent strong vortex centered on the pole, and only a slight shift is observed for cluster 12. In fig. 5a the frequency histogram is shown for the clusters obtained in this study. Cluster 10 is the most frequent state of PV, which is consistent with the current understanding of the nature of PV. Clusters 4-9 and 11

represent shifted PV states of different intensity, which may be estimated by the zonal mean zonal wind at 60°N shown in fig. 5b. In accordance with the recent study [21], there are different shift directions: towards the Atlantic (clusters 7 and 11), towards Eurasia (clusters 4-6 and 8) and North America (cluster 9).

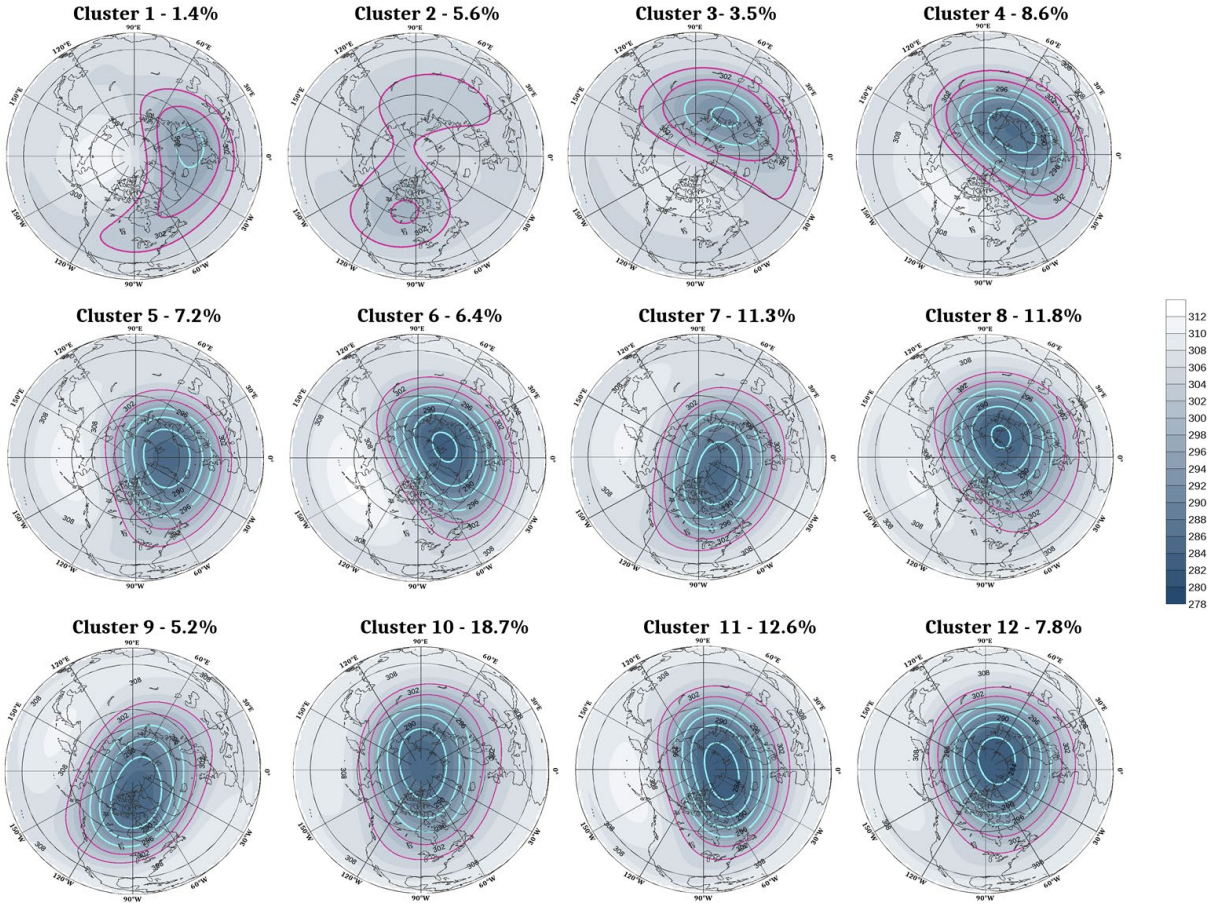
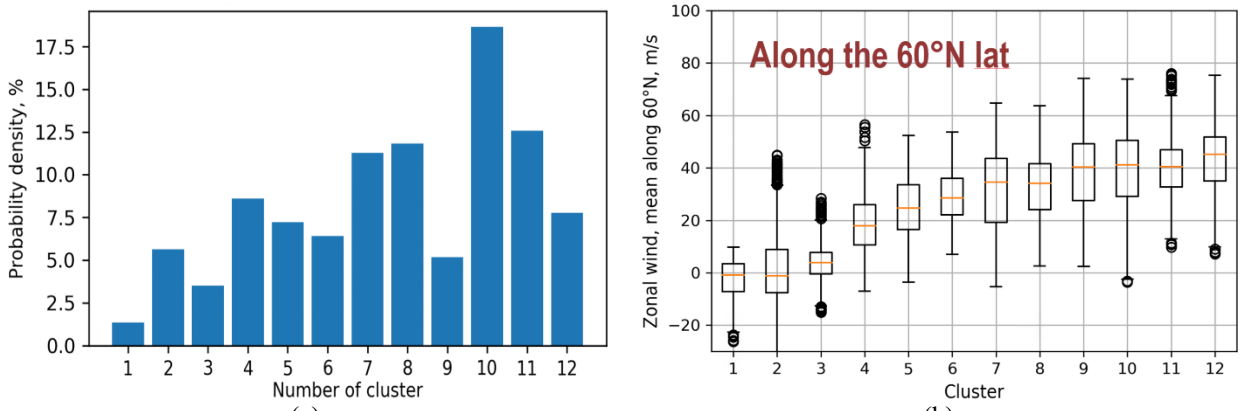


Figure 4: HGT (10^2 m) fields composites for each cluster.



(a) (b)

Figure 5: (a) Histogram of occurrence of PV states associated with each cluster; (b) zonal mean zonal wind along 60°N lat

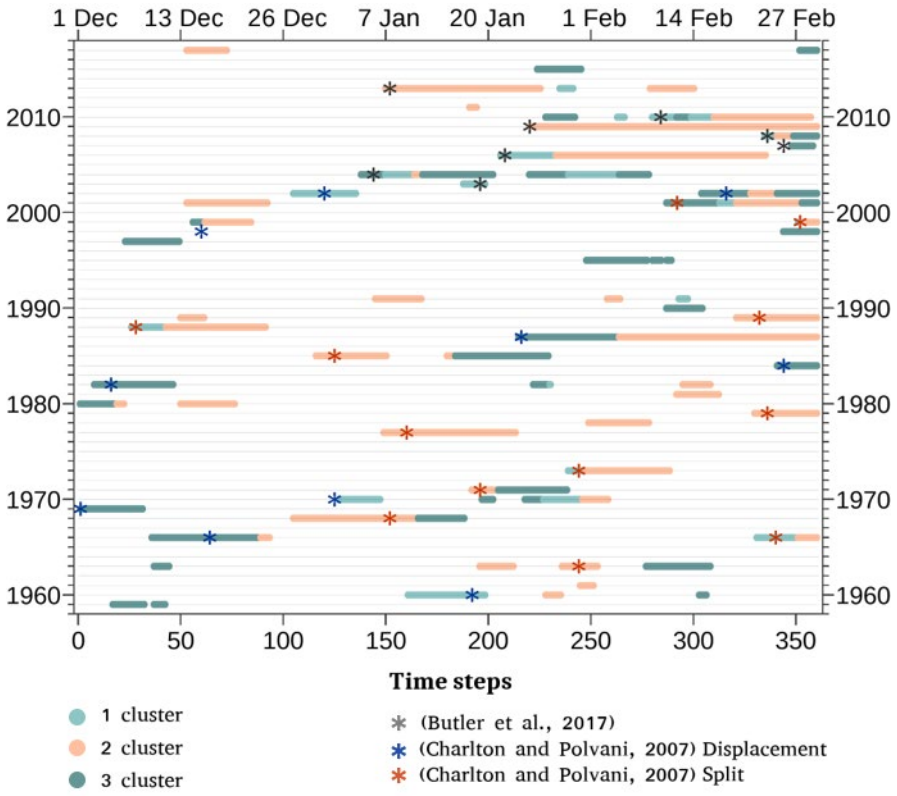


Figure 6: Diagram of occurrences of SSW events based on clustering result (colored segments) and their correspondence to the known SSW events [49,50].

4 Conclusion

We propose a sparse convolutional variational autoencoder as a dimensionality reduction tool. With this model, we extracted valuable features of PV states initially represented by geopotential height and potential vorticity at 10hPa level. Using the representation of reduced dimensionality, we applied the Lance-Williams hierarchical agglomerative clustering with Ward inter-cluster distance definition. This method for the first time is capable of discriminating weak PV states of types "displacement" and "split," which is crucial for the analysis of the stratosphere-troposphere interactions. The proposed method is also capable of classifying stable states of strong PV characterizing by different directions of its center shift. This classification is found to be physically valid and consistent with recent studies. The presented clustering method the first time provides an opportunity to analyze the influence of the strong PV characterized by various shift directions on characteristics of tropospheric circulation.

The proposed clustering method provides an opportunity of researching climatic changes of wintertime stratosphere, which is crucial for improving seasonal forecast skill and assessing the long-term variability of the climate system.

Results of this work can be used as a basis for new stratosphere-troposphere interactions research for existing and future climate scenarios.

Acknowledgments

We thank Joshua Studholme for helpful discussions and suggestions on improving the manuscript. We thank Japan Meteorological Agency (JMA) for making JRA-55 data available.

This research was supported in through computational resources provided by the Shared Facility Center "Data Center of FEB RAS" (Khabarovsk) [51].

This research was supported by the Russian Ministry of Science and Higher Education (agreement № 075-02-2018-189 (14.616.21.0102), project ID RFMEFI61618X0102).

References

1. National Research Council *Assessment of Intraseasonal to Interannual Climate Prediction and Predictability*; 2010; ISBN 978-0-309-15183-2.
2. Butler, A.H.; Polvani, L.M.; Deser, C. Separating the stratospheric and tropospheric pathways of El Niño–Southern Oscillation teleconnections. *Environmental Research Letters* **2014**, *9*, 024014.

3. Polvani, L.M.; Sun, L.; Butler, A.H.; Richter, J.H.; Deser, C. Distinguishing stratospheric sudden warmings from ENSO as key drivers of wintertime climate variability over the North Atlantic and Eurasia. *Journal of Climate* **2017**, *30*, 1959–1969.
4. Kirtman, B.; Pirani, A. The state of the art of seasonal prediction: Outcomes and recommendations from the First World Climate Research Program Workshop on Seasonal Prediction. *Bulletin of the American Meteorological Society* **2009**.
5. Baldwin, M.P.; Dunkerton, T.J. Propagation of the Arctic Oscillation from the stratosphere to the troposphere. *Journal of Geophysical Research: Atmospheres* **1999**, *104*, 30937–30946.
6. Baldwin, M.P.; Dunkerton, T.J. Stratospheric harbingers of anomalous weather regimes. *Science* **2001**, *294*, 581–584.
7. Baldwin, M.P.; Thompson, D.W.; Shuckburgh, E.F.; Norton, W.A.; Gillett, N.P. Weather from the stratosphere? *Science* **2003**, *301*, 317–319.
8. Cohen, J.; Jones, J.; Furtado, J.C.; Tziperman, E. Warm Arctic, cold continents: A common pattern related to Arctic sea ice melt, snow advance, and extreme winter weather. *Oceanography* **2013**, *26*, 150–160.
9. Kidston, J.; Scaife, A.A.; Hardiman, S.C.; Mitchell, D.M.; Butchart, N.; Baldwin, M.P.; Gray, L.J. Stratospheric influence on tropospheric jet streams, storm tracks and surface weather. *Nature Geoscience* **2015**, *8*, 433.
10. Kolstad, E.W.; Breiteig, T.; Scaife, A.A. The association between stratospheric weak polar vortex events and cold air outbreaks in the Northern Hemisphere. *Quarterly Journal of the Royal Meteorological Society* **2010**, *136*, 886–893.
11. Hitchcock, P.; Simpson, I.R. The downward influence of stratospheric sudden warmings. *Journal of the Atmospheric Sciences* **2014**, *71*, 3856–3876.
12. Baldwin, M.P.; Stephenson, D.B.; Thompson, D.W.; Dunkerton, T.J.; Charlton, A.J.; O'neill, A. Stratospheric memory and skill of extended-range weather forecasts. *Science* **2003**, *301*, 636–640.
13. Scaife, A.; Karpechko, A.Y.; Baldwin, M.; Brookshaw, A.; Butler, A.; Eade, R.; Gordon, M.; MacLachlan, C.; Martin, N.; Dunstone, N. Seasonal winter forecasts and the stratosphere. *Atmospheric Science Letters* **2016**, *17*, 51–56.
14. Sigmond, M.; Scinocca, J.; Kharin, V.; Shepherd, T. Enhanced seasonal forecast skill following stratospheric sudden warmings. *Nature Geoscience* **2013**, *6*, 98.
15. Smith, D.M.; Scaife, A.A.; Eade, R.; Knight, J.R. Seasonal to decadal prediction of the winter North Atlantic Oscillation: emerging capability and future prospects. *Quarterly Journal of the Royal Meteorological Society* **2016**, *142*, 611–617.
16. Thompson, D.W.J.; Baldwin, M.P.; Wallace, J.M. Stratospheric Connection to Northern Hemisphere Wintertime Weather: Implications for Prediction. *J. Climate* **2002**, *15*, 1421–1428.
17. Butler, A.H.; Seidel, D.J.; Hardiman, S.C.; Butchart, N.; Birner, T.; Match, A. Defining sudden stratospheric warmings. *Bulletin of the American Meteorological Society* **2015**, *96*, 1913–1928.
18. Cheng, X.; Wallace, J.M. Cluster analysis of the Northern Hemisphere wintertime 500-hPa height field: Spatial patterns. *Journal of the Atmospheric Sciences* **1993**, *50*, 2674–2696.
19. Feldstein, S.B.; Lee, S. Intraseasonal and interdecadal jet shifts in the Northern Hemisphere: The role of warm pool tropical convection and sea ice. *Journal of Climate* **2014**, *27*, 6497–6518.
20. Horton, D.E.; Johnson, N.C.; Singh, D.; Swain, D.L.; Rajaratnam, B.; Diffenbaugh, N.S. Contribution of changes in atmospheric circulation patterns to extreme temperature trends. *Nature* **2015**, *522*, 465.
21. Kretschmer, M.; Coumou, D.; Agel, L.; Barlow, M.; Tziperman, E.; Cohen, J. More-persistent weak stratospheric polar vortex states linked to cold extremes. *Bulletin of the American Meteorological Society* **2018**, *99*, 49–60.
22. Lee, S.; Feldstein, S.B. Detecting ozone-and greenhouse gas–driven wind trends with observational data. *Science* **2013**, *339*, 563–567.
23. Kohonen, T. The self-organizing map. *Proceedings of the IEEE* **1990**, *78*, 1464–1480.
24. Marc'Aurelio Ranzato, F.-J.H.; Boureau, Y.-L.; LeCun, Y. Unsupervised learning of invariant feature hierarchies with applications to object recognition.; 2007; Vol. 127.
25. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536.
26. Kobayashi, S.; Ota, Y.; Harada, Y.; Ebita, A.; Moriya, M.; Onoda, H.; Onogi, K.; Kamahori, H.; Kobayashi, C.; Endo, H. The JRA-55 reanalysis: General specifications and basic characteristics. *Journal of the Meteorological Society of Japan. Ser. II* **2015**, *93*, 5–48.
27. Fujiwara, M.; Wright, J.S.; Manney, G.L.; Gray, L.J.; Anstey, J.; Birner, T.; Davis, S.; Gerber, E.P.; Harvey, V.L.; Hegglin, M.I. Introduction to the SPARC Reanalysis Intercomparison Project (S-RIP) and overview of the reanalysis systems. *Atmospheric Chemistry and Physics* **2017**, *17*, 1417–1452.
28. Oquab, M.; Bottou, L.; Laptev, I.; Sivic, J. Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2014; pp. 1717–1724.

29. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *science* **2006**, *313*, 504–507.
30. Nousi, P.; Tefas, A. Self-supervised autoencoders for clustering and classification. *Evolving Systems* **2018**.
31. Caruana, R. Learning Many Related Tasks at the Same Time with Backpropagation. *Advances in neural information processing systems* **1995**, 8.
32. Collobert, R.; Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the Proceedings of the 25th international conference on Machine learning; ACM, 2008; pp. 160–167.
33. Dauphin, G.M.Y.; Glorot, X.; Rifai, S.; Bengio, Y.; Goodfellow, I.; Lavoie, E.; Muller, X.; Desjardins, G.; Warde-Farley, D.; Vincent, P. Unsupervised and Transfer Learning Challenge: a Deep Learning Approach. In Proceedings of the Proceedings of ICML Workshop on Unsupervised and Transfer Learning; 2012; pp. 97–110.
34. Pan, S.J.; Yang, Q. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* **2010**, *22*, 1345–1359.
35. Pratt, L.Y.; Mostow, J.; Kamm, C.A.; Kamm, A.A. Direct Transfer of Learned Information Among Neural Networks. In Proceedings of the AAAI; 1991; Vol. 91, pp. 584–589.
36. Maclin, R.; Shavlik, J.W. Combining the predictions of multiple classifiers: Using competitive learning to initialize neural networks. In Proceedings of the Proceedings of the 1995 International Joint Conference on AI; Citeseer: Montreal, Quebec, Canada, 1995; pp. 524–531.
37. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural networks* **2015**, *61*, 85–117.
38. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on; Ieee, 2009; pp. 248–255.
39. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]* **2014**.
40. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* **2014**, *15*, 1929–1958.
41. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* **2013**.
42. Krinitskiy, M.; Verezemskaya, P.; Grashchenkov, K.; Tilinina, N.; Gulev, S.; Lazzara, M.; Krinitskiy, M.; Verezemskaya, P.; Grashchenkov, K.; Tilinina, N.; et al. Deep Convolutional Neural Networks Capabilities for Binary Classification of Polar Mesocyclones in Satellite Mosaics. *Atmosphere* **2018**, *9*, 426.
43. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multiscale structural similarity for image quality assessment. In Proceedings of the The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003; 2003; Vol. 2, pp. 1398-1402 Vol.2.
44. Milligan, G.W. Ultrametric hierarchical clustering algorithms. *Psychometrika* **1979**, *44*, 343–346.
45. Murtagh, F.; Contreras, P. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2012**, *2*, 86–97.
46. Murtagh, F.; Contreras, P. Algorithms for hierarchical clustering: an overview, II. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2017**, *7*, e1219.
47. Jr, J.H.W. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* **1963**, *58*, 236–244.
48. Mandel, I.D. Klasternyj analiz. *Finansy i statistika* **1988**, *176*.
49. Charlton, A.J.; Polvani, L.M. A new look at stratospheric sudden warmings. Part I: Climatology and modeling benchmarks. *Journal of Climate* **2007**, *20*, 449–469.
50. Butler, A.H.; Sjoberg, J.P.; Seidel, D.J.; Rosenlof, K.H. A sudden stratospheric warming compendium. *Earth System Science Data* **2017**, *9*.
51. Sorokin, A.A.; Makogonov, S.V.; Korolev, S.P. The Information Infrastructure for Collective Scientific Work in the Far East of Russia. *Sci. Tech. Inf. Proc.* **2017**, *44*, 302–304.