TOWARDS AN IMPROVEMENT OF THE FORECAST OF WIND RESOURCES IN EUROPE: APPLICATION OF UNSUPERVISED MACHINE LEARNING ON FUTURE PROJECTIONS OF POLAR VORTEX

María Rodríguez Montes

Computer Engineering Ph.D. student
Complutense University of Madrid
marodr51@ucm.es

Blanca Ayarzagüena

Dept. Earth Physics and Astrophysics
Complutense University of Madrid
bayarzag@ucm.es

María Guijarro Mata-García

Dept. Computer Architecture and Automatic Control
Complutense University of Madrid
mguijarro@ucm.es

ABSTRACT

Studying the impact of climate change on wintertime polar stratosphere is of particular relevance not only for climate knowledge but also for tropospheric projections. Machine learning provides a way to extract information from different climate models and combine the data in a such way that patterns are clearer, so predictions can be inducted. The methods used in this study have been region growing algorithm, K-means and combination of predictions. The final results show three clusters of response trends for the intensity and location of the stratospheric polar night jet. The prediction shows an increase in zonal wind intensity over the 2xCO2 and 4xCO2 concentration points and a decrease in the related latitude. Our methods can be extended for more climate models and simulation periods, and will allow not only to map the behavior of the polar night jet, but other stratospheric and tropospheric features and interactions between them.

Keywords: forecasting, K-means, polar vortex, polar night jet, region growing

1 INTRODUCTION

Climate science is one of the disciplines with a high potential to apply artificial intelligence (AI) and data mining. It typically deals with a large amount of data even in the simplest studies. It also implies the use of large amount of computer resources. For instance, climate predictions are performed thanks to simulations of global climate models. In most of the cases, these climate models involve the numerical solution of millions of non-linear mathematical equations that reproduce the processes and interactions of the climate system components (atmosphere, ocean, ice, vegetation and land). As a result, running a single climate simulation may last several months, particularly when simulating several decades. Thus, machine learning techniques can be useful in climate as can speed up analyses or extract relevant information from different climate models as suggested by Huntingford et al. (2019). However, only just recently AI methodology has been applied to climate analyses.

One of the first attempts of applying data mining to climate simulations has been the use of a Bayesian approach to develop forecasts of oceanic and atmospheric variables (Luo et al. 2007). Through cross validation of these forecasts they could indicate an improvement in both, deterministic and probabilistic forecast skills. More recently, Monteleoni et al. (2011) used an hierarchical learner algorithm based on a set of generalized Hidden Markov Models, HMM, to perform predictions based on different climate model simulations.

Totz et al. (2017) have focused their research on future climate over Europe, especially the Mediterranean region where climate models project a reduction in winter rainfall and a very pronounced increase in summertime heat waves in the future. They proposed a method to predict precipitation anomalies in winter inspired in a new cluster-based empirical forecast algorithm.

Considering the large amount of data provided by climate models, neural networks have also been shown to offer very satisfactory solutions as powerful tools for classifying, identifying, and predicting patterns in climate and environmental data. Chattopadhyay et al. (2019) proposed a method based on the deep learning pattern recognition technique by capsule neural networks, CapsNets, and the impact-based automatic labeling strategy applied to large-scale circulation patterns in the middle troposphere. To address these challenges Chattopadhyay et al. (2020) continued their research and proposed an effective auto-labeling strategy based on using an unsupervised clustering algorithm and evaluating the performance of convolutional neural networks in re-identifying and predicting these clusters. To prove this efficiency they use this approach to label thousands of daily large-scale weather patterns over North America in the outputs of a climate model, obtaining an accuracy of 90%.

Given the relatively recent application of AI and data mining to climate sciences, analyses have mainly focused on the first atmospheric layer (the troposphere). However, while the troposphere accounts for the 85% of the total mass of the atmosphere, the stratosphere (the next atmospheric layer) also plays an active role in climate (Baldwin et al. 2001). Indeed, changes in the intensity of the stratospheric polar vortex (a strong circumpolar circulation present from Autumn to Spring) have been linked to changes in precipitation and near-surface circulation over Europe in the following two months. Actually, the effects of these extreme events on European climate are so important that they can even provide predictability of wind electricity generation in that area (Beerli et al. 2017). Despite the key role of the stratosphere, only a few studies such as Kretschmer et al. (2016) and Kretschmer et al. (2017) have applied data mining and causal effect network to investigate it. Kretschmer et al. (2016) analyzed the effects of the polar stratospheric variability on near surface climate, whereas Kretschmer et al. (2017) predict extreme stratospheric events in a time scale of weeks. However, applying AI might also reduce some uncertainties on stratosphere-troposphere climate. For instance, the most recent climate models do not even agree on the sign of the polar stratospheric response to increasing greenhouse gases concentrations (Ayarzagüena et al. 2020). This in turn adds uncertainty in the future tropospheric projections over Europe. Applying AI techniques to different model simulations might help to provide a better picture of future stratosphere-troposphere projections.

In this study, we aim to obtain, for the first time, a prediction of the effects of increasing CO2 concentrations on stratospheric polar vortex by applying data mining to an ensemble of simulations of the most recent version of climate models. The paper is organized as follows: section 2 presents a description of the calculation methods applied throughout the document; section 3 describes the climate data including variables and models; section 4 presents the results that are obtained; finally, the conclusions of the study are presented in section 5.

2 METHODS

Here, we present a brief introduction of the full process in time series forecasting that has been followed in this study:

1. Obtaining the initial database of F features ($i = 1...F$), each of them containing S samples ($j = 1...S$). Each sample in the database has three time points. Therefore, the total number of data points is $F \times S \times 3$. Each of the time series will be referred to as y_{ij}. The calculation of the climate database studied in this document is described in section 3.3.
2. Validation of the database. The validation of the specific climate database in this project is described in section 3.3.
3. Calculation of the trend features vector for each time series y_{ij}. The trend features vector for each time series is $\overline{f}_{ij}$. The method for calculating trend features is described in section 2.1.
4. K-means clusters applied individually to the group of trend features $\overline{f}_i$ for obtaining M clusters. The general clustering process is described in section 2.2, and, for this specific use case, in section 4.1.
5. Clustering algorithm validation. The specific validation for this climate use case is described in section 4.1.
6. For each variable in the database, calculation of M predictions, corresponding to each cluster. This method is described in 2.3.
7. Combination of the M predictions into one with weighted averaging. This method is described in section 2.3 and, specifically for the climate database in section 4.2.

2.1 Trend Features Vector

The trend of each of the samples y_{ij} in the database is described with a 4-component vector:

$$\overline{f}_{ij} = [|m_1|, |m_2|, sign(m_1), sign(m_2)] \tag{1}$$

where m_1 means slope of the line joining initial point and second measurement, m_2 means slope of the line joining the second and final or third measurement.

Both, the value of the slope and its sign are important when analyzing patterns of behavior. The absolute values of the slopes and their signs have been separated into the features vector in order not to have information canceled out in the K-means algorithm, described in section 2.2. For this reason, if the vector in equation (1) has only two components ($[m_1, m_2]$), some information could disappear through the K-means algorithm. This is the case in time series with opposite signs and very close values of m_1. To avoid the loss of information, our algorithm separates sign and value of the trend information.

The trend features database is pre-processed before introducing it in the clustering algorithm. Each trend feature is scaled individually so that its value is between zero and one. Also, samples whose m_1 ad m_2 are both below 5% the value of the maximum for each group of trend features $\overline{f}_i$ are eliminated from the database; this percentage is considered a negligible response in the use case of this document.

2.2 K-means Algorithm

K-means is an unsupervised machine learning algorithm used for clustering (Yuan et al. 2019). In this case, it is going to be used for clustering the group of features $\overline{f}_i = [|m_1|, |m_2|, sign(m_1), sign(m_2)]$ shown in equation (1).

The inputs to K-means are a group of data and the number of M centroids that are desired as a result. The output is M centroids and labels for each data point; each label is the centroid to which that data point is assigned.

The algorithm consists of the following steps: the first one is initialization. The second step consists of performing the following actions until convergence: data assignment, where each data point is assigned to

the nearest centroid, by obtaining the minimum of the squared Euclidean distance; and centroid update, where centroids are recomputed. New centroids are obtained by taking the mean of all data points assigned to each centroid in the previous step.

The K-means method implemented in the calculations comes from the Python library scikit-learn (Pedregosa et al. 2011).

2.3 Forecasting and Combination of Predictions

Once the cluster centers have been calculated by the K-means algorithm, information is available about slope between first, second and third point, and their respective signs, from the each of the four component center $[|m_1|, |m_2|, sign(m_1), sign(m_2)]$. This information can be used to build two straight lines that predict a time series that starts at an initial condition. The predicted time series are defined by equations (2). In the use case of this document, y will be a climate variable, and t will be measured in years.

$$\begin{aligned} y(t_2) &= y(t_1) + m_{1,cluster} \cdot (t_2 - t_1) \\ y(t_3) &= y(t_2) + m_{2,cluster} \cdot (t_3 - t_2) \end{aligned} \tag{2}$$

Equations (2) give as many predictions as the number of cluster centers for each database feature. There are several ways to combine these predictions, so that in the end there is only one prediction. Some methods are uniform averaging, weighted averaging and Naïve classifiers (Trivedi et al. 2015). For this study, average weighing has been selected.

3 DATA

3.1 Climate Models Data

In this study we use the output of simulations performed by state-of-the-art climate models participating in the recent Coupled Model Intercomparison Project, Phase 6 (CMIP6). These climate models aim to provide the most detailed representation of the climate system since they include as many of the Earth system processes as possible. In the CMIP6 initiative the main international climate modelling centers have contributed with simulations following the specifications provided by the CMIP Panel regarding forcings such as greenhouse gases concentrations or solar forcings (Eyring et al. 2016). These are the simulations that are currently being used to elaborate the next report of the Intergovernmental Panel on Climate Change (IPCC).

More specifically, here we use monthly data of zonal wind (u) of the 1pctCO2 simulation of the CMIP6 models indicated in Table 1. The 1pctCO2 run extends 150 years with a gradual increase of the CO2 concentration at a rate of 1% per year that starts at the pre-industrial level (year 1850, 284.32 ppm). In this simulation the average mean concentration of CO2 in our recent past period (1958-2010) is reached after approximately 20 years and corresponds to 284.32 pm, approximately 1.3 times the pre-industrial concentrations (Meinshausen et al. 2017). Actually, these concentrations of CO2 for the recent past period range from 315.34 ppm in 1958 to 388.72 ppm in 2010. In the 1pctCO2 simulation those values correspond to years 11 and 31, respectively, so we will consider these 21 years to characterize the first time step of our analysis (denoted by 1.3xCO2). The selection of this time step enables the validation of models. In addition, in the same run the CO2 concentrations of the pre-industrial era (284.32 ppm) are doubled after approximately 70 years and quadrupled after approximately 140 years. We consider these two other time steps in our analysis, by selecting the 10 years surrounding year 70 and year 140 (hereafter denoted as 2xCO2 and 4xCO2, respectively).

Table 1: List of models included in the analysis

Models	Model reference
CanESM5	Swart et al. (2019b), Swart et al. (2019a)
CESM2	Danabasoglu (2019a), Danabasoglu et al. (2020)
CESM2-WACCM	Danabasoglu (2019b), Gettelman et al. (2019)
CNRM-CM6-1	Voldoire (2018), Voldoire et al. (2019)
CNRM-ESM2-1	Seferian (2018), Séférian et al. (2019)
GFDL-CM4	Guo et al. (2018), Held et al. (2019)
GISS-E2-2-G	NASA/GISS (2018)
HadGEM3-GC31-LL	Roberts (2017), Williams et al. (2018)
INM-CM5-0	Volodin et al. (2017)
IPSL-CM6A-LR	Boucher et al. (2018)
MIROC6	Tatebe et al. (2018), Tatebe et al. (2019)
MRI-ESM2-0	Yukimoto et al. (2019b), Yukimoto et al. (2019a)
UKESM1-0-LL	Tang et al. (2019), Kuhlbrodt et al. (2018)

3.2 Reanalysis Data

Apart from model simulations, monthly mean data of zonal wind of the JRA-55 reanalysis (Kobayashi et al. 2015) for the period 1958-2010 is used. Although reanalysis data are not exactly direct observations, they can be considered as the real world of the last decades. They are derived from the assimilation of observations of different sources that are then ingested by a model that produce an homogeneous data set. The use of this data here has two purposes. First, it allows the validation of models by the comparison of the simulation of the polar night jet in models in 1.3xCO2 step with reanalysis results. Secondly, as indicated in the following sections, it constitutes the initial value to produce the predictions of the polar night jet state under increasing CO2 concentrations.

3.3 Polar Night Jet Features

As indicated in the Introduction, stratospheric polar vortex is the main circulation structure in the polar stratosphere in winter. It consists of a strong cyclonic circulation located over the polar cap. The core of this cyclonic circulation or edge of the polar vortex is called the polar night jet (PNJ). The main characteristics of the PNJ are $u_{c,PNJ}$ and ϕ_{PNJ}. $u_{c,PNJ}$ corresponds to the climatological intensity of the PNJ and is computed as the average of u_c in the PNJ region in December-January-February) at 10hPa (the middle stratosphere), in the relevant time points from the 1pctxCO2 simulation. ϕ_{PNJ} is the average latitude of the PNJ region.

Before calculating these magnitudes, the PNJ region is identified for each point in the time series and each climate model of table 1. The PNJ region has been identified with a region growing algorithm which starts at a seed value. In this case, the seed consists of the latitude and longitude with the maximum u_c. In the first step of the region growing algorithm, the 8 neighbors of the seed point are checked for their u_c. They are appended to the PNJ zone if their u_c value is greater than a threshold value. The threshold value has been adapted to each climate model, and it runs from 65% to 75% of the maximum u_c. The same steps are followed for each appended point, until reaching an edge where the threshold condition is no longer met.

After the identification of the PNJ zone, $u_{c,PNJ}$ and ϕ_{PNJ} are calculated on this region as a weighted average of the values. $u_{c,PNJ}$ is calculated by weighing u_c with the cosine of the latitude at each point in the region, whereas ϕ_{PNJ} is calculated by weighing the latitude of each point in the PNJ with u_c at that point.

An example of the results of region growing algorithm is shown in figures 1 and 2. Figure 1 shows the distribution of u_c over the Earth surface at 10 hPa, for the CESM2-WACCM model at the 1.3xCO2 point of the time analysis. As was previously mentioned, u_c refers to the climatology of u in the winter months. Figure 2 shows the polar night jet region as detected by the region growing algorithm that has been applied to the data in figure 1.

The calculated data for the PNJ database is presented in figure 3. The left figure of 3 shows $u_{c,PNJ}$ for the three time steps of the analysis 1.3xCO2, 2xCO2 and 4xCO2, whereas the figure on the right shows the approximate center of the polar night jet ϕ_{PNJ} for the same three time points. The corresponding reanalysis data points from the simulation, which will be taken as initial conditions for predictions, are also shown in these figures. The context of the reanalysis data was described in section 3.2.

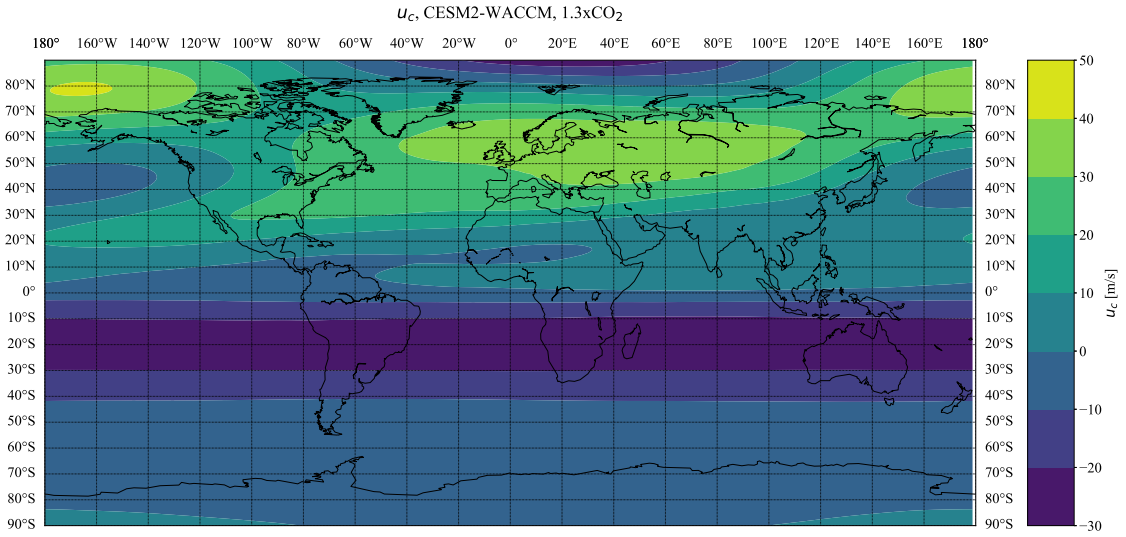

Figure 1: Climatology of u (m/s) at 10hPa in winter months for CESM2-WACCM model in 1.3xCO2 data point.

The PNJ data shown in figure 3 has been validated by comparing their trends with the calculation of the maximum of u_c's zonal mean, and its associated latitude. These two magnitudes are also representative of the average intensity and location of the PNJ. The results are presented in the following table 2, which shows the relative errors between the sign of each time segment (s_1 and s_2) of each atmospheric feature in this document, as calculated with the region growing approach, and the approach described at the beginning of this paragraph. Most relative errors are 0%: the signs of each of the two time segments are, in general, similar. Therefore, the region growing algorithm as it has been applied here is considered validated, except for future potential improvements on the segments whose signs do not match.

4 CLUSTERING AND PREDICTION RESULTS

4.1 K-means Clustering Results

Before launching the K-means calculations, the number of clusters has to be set. There are several methods for doing this, but in this case the optimal number of centroids for launching K-means has been chosen by analysing the curve of inertia indicator versus the number of clusters. The inertia indicator is the sum of squared distances between the samples and the closest cluster center. The optimal number of clusters has been chosen as the point where this curve begins to linearly decrease, a method otherwise known as elbow

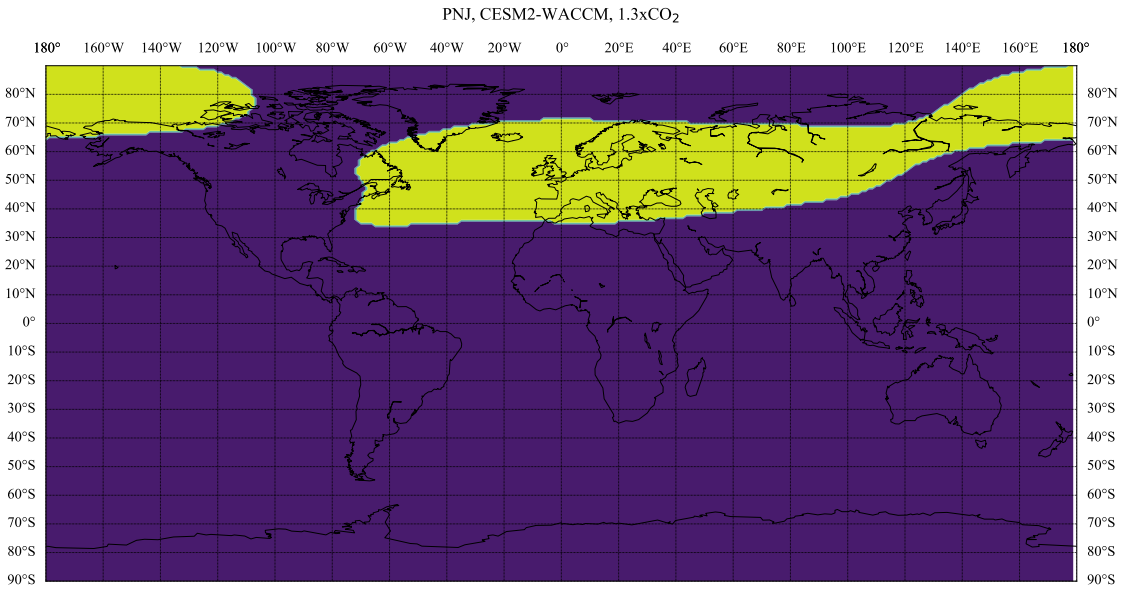

Figure 2: PNJ region derived from figure 1 for CESM2-WACCM model in 1.3xCO2 data point.

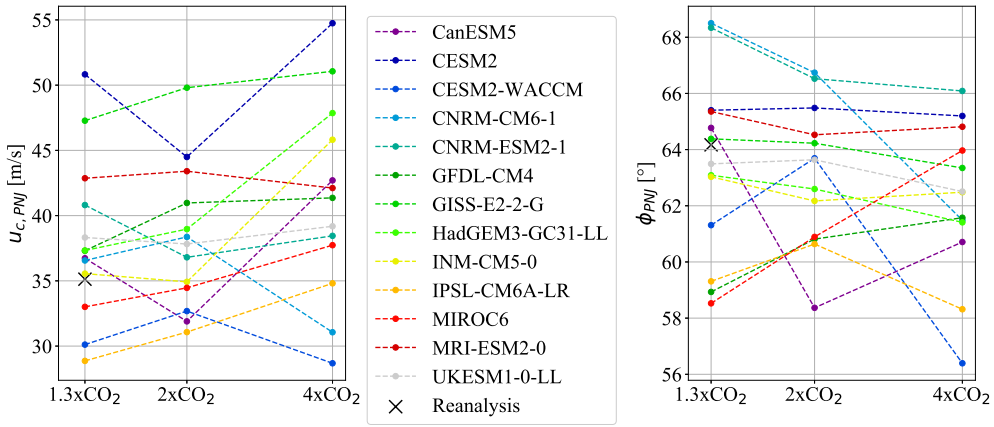

Figure 3: $u_{c,PNJ}$ and ϕ_{PNJ} for three CO2 concentrations over time in each model.

method (Yuan et al. 2019). The analysed curve for $u_{c,PNJ}$ is shown in figure 4, where M=3 clusters have been selected as an optimal number. The criterion has been the start of the decrease of the second derivative of the curve at M =3, which in this case, means the curve is beginning to look like a straight line.

The clusters produced by applying the K-means algorithm to the features vector of the climate database are shown in tables 3 and 4. The information in these tables means that, for each variable $u_{c,PNJ}$ and ϕ_{PNJ} in the database, there are three clusters. The models assigned to each cluster are considered to have the same behavior according to the K-means algorithm and the trend features vector given in equation (1).

As mentioned in section 2.2, the information that we get from the cluster centers is the slope between first and second time point, slope between second and third time point and their respective signs ($\overline{f}_i = [|m_1|, |m_2|, sign(m_1), sign(m_2)]$). The cluster centers are the inputs to the prediction equations (2), which will be combined afterwards with weighted averaging.

Table 2: Relative error of sign of trend of two time segments for two approaches for calculation of $u_{c,PNJ}$ and ϕ_{PNJ}.

Model / Relative error [%]	ϕ_{PNJ} s_1	ϕ_{PNJ} s_2	$u_{c,PNJ}$ s_1	$u_{c,PNJ}$ s_2
CanESM5	0	-100	0	0
CESM2	0	0	0	0
CESM2-WACCM	0	0	0	0
CNRM-CM6-1	0	0	0	0
CNRM-ESM2-1	0	0	0	0
GFDL-CM4	0	-100	-100	0
GISS-E2-2-G	0	-100	0	-100
HadGEM3-GC31-LL	0	0	0	0
INM-CM5-0	-100	0	0	0
IPSL-CM6A-LR	0	0	0	0
MIROC6	0	0	0	-100
MRI-ESM2-0	0	0	0	0
UKESM1-0-LL	0	0	0	0

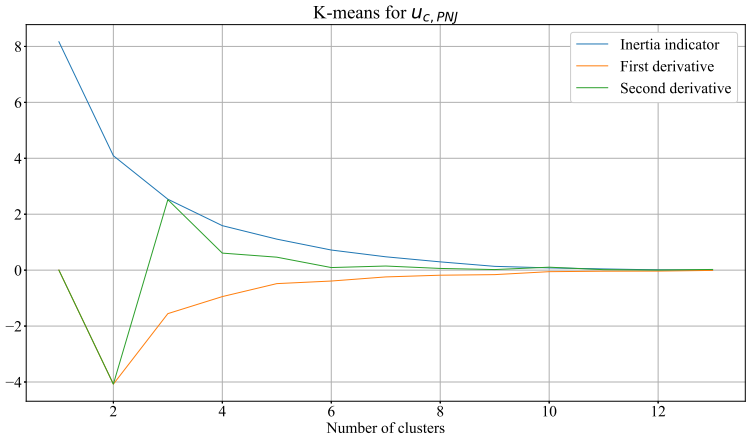

Figure 4: Inertia indicator analysis for K-means clustering of $u_{c,PNJ}$ trend features.

Table 3: Models clusters for $u_{c,PNJ}$

Cluster 1	Cluster 2	Cluster 3
CanESM5	GISS-E2-2-G	CESM2-WACCM
CESM2	HadGEM3-GC31-LL	CNRM-CM6-1
CNRM-ESM2-1	IPSL-CM6A-LR	GFDL-CM4
INM-CM5-0	MIROC6	MRI-ESM2-0
UKESM1-0-LL		

The clustering algorithm has been validated by applying the algorithm to the $u_{c,PNJ}$ trend between the piControl simulation (pre-industrial control simulation, Eyring et al. (2016), initial conditions of the 1pctCO2 run) and the last time point analysed in this document (4xCO2), and comparing it with the results presented in Ayarzagüena et al. (2020), where the trends for the same climate variable for the period between piControl

Table 4: Models clusters for ϕ_{PNJ}

Cluster 1	Cluster 2	Cluster 3
CanESM5	CNRM-CM6-1	CESM2-WACCM
GFDL-CM4	CNRM-ESM2-1	IPSL-CM6A-LR
MIROC6	GISS-E2-2-G	
	HadGEM3-GC31-LL	
	INM-CM5-0	
	MRI-ESM2-0	
	UKESM1-0-LL	

and abrupt4xCO2 simulation (Eyring et al. 2016) are discussed. In general, the trends have been observed to be similar. There is a difference in the trends of GFDL-CM4 and UKESM1-0-LL: one analysis shows no response versus the other, which shows a response. This is possibly due to the difference in calculation methods. Note that the clusters given in table 5 are different from those presented in the previous tables 3 and 4.

Table 5: Trends between piControl and 4xCO2 (1pctCO2) or abrupt4xCO2 simulation periods.

Model	Trend	Trend (Ayarzagüena et al. 2020)
CanESM5	Cluster1, ↑	↑
CESM2	Cluster1, ↑	↑
GFDL-CM4	Cluster 1, ↑	No response
GISS-E2-2-G	Cluster1, ↑	↑
IPSL-CM6A-LR	Cluster 1, ↑	↑
MIROC6	Cluster1, ↑	↑
HadGEM3-GC31-LL	Cluster 2, ↑	↑
INM-CM5-0	Cluster 2, ↑	↑
CESM2-WACCM	Cluster 3, ↓	↓
CNRM-CM6-1	Cluster 3, ↓	-
CNRM-ESM2-1	Cluster 3, ↓	↓
MRI-ESM2-0	Cluster 4, no response	No response
UKESM1-0-LL	Cluster 4, no response	↑

4.2 Application of Prediction Algorithm

Referring to the combination of predictions, a weight has been given to each of the cluster centers obtained in section 4.1. This weight has been calculated based on the average value of the different variables for the models in each cluster at point 1.3xCO2, and how this result compares with the reanalysis values, by means of a Euclidean distance. A weighted average is then calculated with the three predictions for both variables separately.

The prediction results obtained after applying the full process described in the introduction of section 2, for the two studied atmospheric features, are shown in figure 5. We observe that the climate variable $u_{c,PNJ}$ increases from its initial point to a final position which is higher than the initial condition, whereas the climate variable ϕ_{PNJ} decreases in time. Thus, in general, the PNJ would become stronger and shift equatorward under increasing CO2 concentrations. The PNJ shift would indicate that the polar vortex would become larger, which agrees well with its intensification.

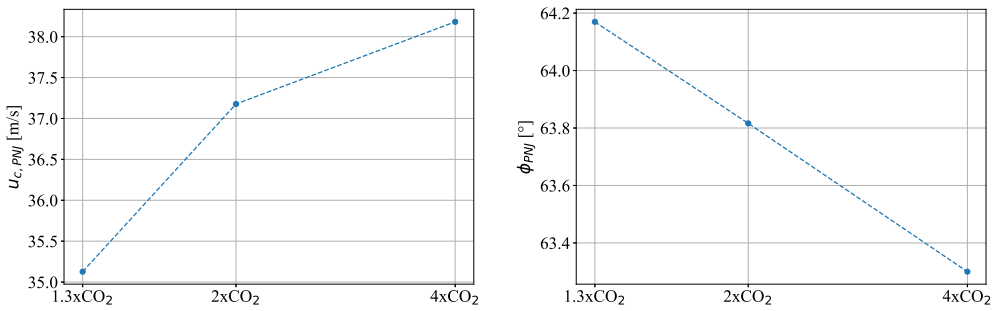

Figure 5: Prediction of $u_{c,PNJ}$ and ϕ_{PNJ} at 2xCO2 and 4xCO2.

5 CONCLUSIONS

In this paper, we have presented a method to obtain future projections of stratospheric polar night jet based on mining a group of climate models at different atmospheric carbon levels. The goodness of fit of the method has been proven by validating the climate database and the clustering algorithm with alternative approaches commonly used in climate physics. The results of the predictions indicate that there will be a decrease in the center latitude of the polar night jet at 2xCO2 and 4xCO2 concentrations, and that the average zonal wind climatology on the polar night jet will also suffer an increase in these two CO2 concentrations. This would indicate a stronger polar vortex in the future, which could also then impact on surface weather and wind resources over Europe.

ACKNOWLEDGEMENTS

This work was partially supported by the Spanish Ministry of Science, Innovation and Universities under Project number RTI2018-094902-B-C21. We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP6. We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP6 and ESGF. CMIP6 data are allocated at the ESGF archive (https://esgf-node.llnl.gov/projects/ cmip6/). The Japanese 55-year Reanalysis (JRA-55) project was carried out by the Japan Meteorological Agency (JMA). JRA-55 data were accessed through NCAR- UCAR Research Data Archive (https://rda.ucar.edu).

REFERENCES

Ayarzagüena et al. 2020. “Uncertainty in the response of sudden stratospheric warmings and stratosphere-troposphere coupling to quadrupled CO2 concentrations in CMIP6 models”. *Journal of Geophysical Research: Atmospheres* vol. 125. doi: 10.1029/2019JD032345.

Baldwin et al. 2001. “Stratospheric Harbingers of Anomalous Weather Regimes”. *Science* vol. 294 (5542), pp. 581–584. American Association for the Advancement of Science, doi: 10.1126/science.1063315.

Beerli et al. 2017. “Does the lower stratosphere provide predictability for month-ahead wind electricity generation in Europe?”. *Quarterly Journal of the Royal Meteorological Society* vol. 143 (709), pp. 3025–3036.

Boucher et al. 2018. “IPSL IPSL-CM6A-LR model output prepared for CMIP6 CMIP”. Earth System Grid Federation, doi: 10.22033/ESGF/CMIP6.1534.

Chattopadhyay et al. 2019. “Analog forecasting of extreme-causing weather patterns using deep learning”. *arXiv preprint arXiv:1907.11617*.

Chattopadhyay et al. 2020. "Predicting clustered weather patterns: A test case for applications of convolutional neural networks to spatio-temporal climate data". *Scientific Reports* vol. 10 (1), pp. 1–13. Nature Publishing Group.

Danabasoglu, G. 2019a. "NCAR CESM2 model output prepared for CMIP6 CMIP 1pctCO2". Earth System Grid Federation, doi: 10.22033/ESGF/CMIP6.7497.

Danabasoglu, G. 2019b. "NCAR CESM2-WACCM model output prepared for CMIP6 CMIP". Earth System Grid Federation, doi: 10.22033/ESGF/CMIP6.10024.

Danabasoglu et al. 2020. "The Community Earth System Model Version 2 (CESM2)". *Journal of Advances in Modeling Earth Systems* vol. 12 (2), pp. e2019MS001916. doi: 10.1029/2019MS001916.

Eyring et al. 2016. "Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organizations". *Geoscientific Model Development* vol. 9, pp. 1937–1958.

Gettelman et al. 2019. "The Whole Atmosphere Community Climate Model Version 6 (WACCM6)". *Journal of Geophysical Research: Atmospheres* vol. 124 (23), pp. 12380–12403. doi: 10.1029/2019JD030943.

Guo et al. 2018. "NOAA-GFDL GFDL-CM4 model output". Earth System Grid Federation, doi: 10.22033/ESGF/CMIP6.1402.

Held et al. 2019. "Structure and Performance of GFDL's CM4.0 Climate Model". *Journal of Advances in Modeling Earth Systems* vol. 11 (11), pp. 3691–3727. doi: 10.1029/2019MS001829.

Huntingford et al. 2019. "Machine learning and artificial intelligence to aid climate change research and preparedness". *Environmental Research Letters* vol. 14 (12), pp. 124007.

Kobayashi et al. 2015. "The JRA-55 reanalysis: General specifications and basic characteristics". *Journal of the Meteorological Society of Japan* vol. 93, pp. 5–48.

Kretschmer et al. 2016. "Using Causal Effect Networks to Analyze Different Arctic Drivers of Midlatitude Winter Circulation". *Journal of Climate* vol. 29 (11), pp. 4069–4081.

Kretschmer et al. 2017. "Early prediction of extreme stratospheric polar vortex states based on causal precursors". *Geophysical Research Letters* vol. 44 (16), pp. 8592–8600.

Kuhlbrodt et al. 2018. "The Low-Resolution Version of HadGEM3 GC3.1: Development and Evaluation for Global Climate". *Journal of Advances in Modeling Earth Systems* vol. 10 (11), pp. 2865–2888. doi: 10.1029/2018MS001370.

Luo et al. 2007. "Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions". *Journal of Geophysical Research: Atmospheres* vol. 112 (D10).

Meinshausen et al. 2017. "Historical greenhouse gas concentrations for climate modelling (CMIP6)". *Geoscientific Model Development* vol. 10, pp. 2057–2116.

Monteleoni et al. 2011. "Tracking climate models". *Statistical Analysis and Data Mining: The ASA Data Science Journal* vol. 4 (4), pp. 372–392. Wiley Online Library.

NASA/GISS 2018. "NASA-GISS GISS-E2.1G model output prepared for CMIP6 CMIP piControl". Earth System Grid Federation, doi: 10.22033/ESGF/CMIP6.7380.

Pedregosa et al. 2011. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research* vol. 12, pp. 2825–2830.

Roberts, M. 2017. "MOHC HadGEM3-GC31-LL model output prepared for CMIP6 HighResMIP". Earth System Grid Federation, doi: 10.22033/ESGF/CMIP6.1901.

Seferian, R. 2018. "CNRM-CERFACS CNRM-ESM2-1 model output prepared for CMIP6 CMIP for experiment piControl-spinup". Earth System Grid Federation, doi: 10.22033/ESGF/CMIP6.4169.

Swart et al. 2019a. "The Canadian Earth System Model version 5 (CanESM5.0.3)". *Geoscientific Model Development* vol. 12 (11), pp. 4823–4873. doi: 10.5194/gmd-12-4823-2019.

Swart et al. 2019b. "CCCma CanESM5 model output prepared for CMIP6 CMIP 1pctCO2". Earth System Grid Federation, doi: 10.22033/ESGF/CMIP6.3151.

Séférian et al. 2019. "Evaluation of CNRM Earth System Model, CNRM-ESM2-1: Role of Earth System Processes in Present-Day and Future Climate". *Journal of Advances in Modeling Earth Systems* vol. 11 (12), pp. 4182–4227. doi: 10.1029/2019MS001791.

Tang et al. 2019. "MOHC UKESM1.0-LL model output prepared for CMIP6 CMIP 1pctCO2". Earth System Grid Federation, doi: 10.22033/ESGF/CMIP6.5792.

Tatebe et al. 2018. "MIROC MIROC6 model output prepared for CMIP6 CMIP 1pctCO2". Earth System Grid Federation, doi: 10.22033/ESGF/CMIP6.5371.

Tatebe et al. 2019. "Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6". *Geoscientific Model Development* vol. 12 (7), pp. 2727–2765. doi: 10.5194/gmd-12-2727-2019.

Totz et al. 2017. "Winter precipitation forecast in the European and Mediterranean regions using cluster analysis". *Geophysical Research Letters* vol. 44 (24), pp. 12–418. Wiley Online Library.

Trivedi et al. 2015. "The Utility of Clustering in Prediction Tasks". *CoRR* vol. abs/1509.06163.

Voldoire, A. 2018. "CMIP6 simulations of the CNRM-CERFACS based on CNRM-CM6-1 model for CMIP experiment 1pctCO2". Earth System Grid Federation, doi: 10.22033/ESGF/CMIP6.3712.

Voldoire et al. 2019. "Evaluation of CMIP6 DECK Experiments With CNRM-CM6-1". *Journal of Advances in Modeling Earth Systems* vol. 11 (7), pp. 2177–2213. doi: 10.1029/2019MS001683.

Volodin et al. 2017. "Simulation of the present-day climate with the climate model INMCM5". *Climate Dynamics* vol. 49, pp. 3715–3734.

Williams et al. 2018. "The Met Office Global Coupled Model 3.0 and 3.1 (GC3.0 and GC3.1) Configurations". *Journal of Advances in Modeling Earth Systems* vol. 10 (2), pp. 357–380. doi: 10.1002/2017MS001115.

Yuan et al. 2019, Jun. "Research on K-Value Selection Method of K-Means Clustering Algorithm". *J* vol. 2 (2), pp. 226–235. MDPI AG, doi: 10.3390/j2020016.

Yukimoto et al. 2019a. "The Meteorological Research Institute Earth System Model Version 2.0, MRI-ESM2.0: Description and Basic Evaluation of the Physical Component". *Journal of the Meteorological Society of Japan. Ser. II* vol. 97 (5), pp. 931–965. doi: 10.2151/jmsj.2019-051.

Yukimoto et al. 2019b. "MRI MRI-ESM2.0 model output prepared for CMIP6 CMIP". Earth System Grid Federation, doi: 10.22033/ESGF/CMIP6.621.

AUTHOR BIOGRAPHIES

MARÍA RODRÍGUEZ is an Aerospace Engineer and a Computer Engineering Ph.D. student at Complutense University of Madrid. Her research interests include computer and climate sciences.

BLANCA AYARZAGÜENA is a Lecturer at the University Complutense of Madrid. She holds a Ph.D in Atmospheric Sciences. Her research interests include climate change and stratospheric dynamics.

MARÍA GUIJARRO is an Associate Professor in the University Complutense of Madrid. She holds a Ph.D. in AI based on Computer Vision Algorithms. Her research interests include modeling and simulation.