Large ensemble assessment of the Arctic stratospheric polar vortex

Ales Kuchar[1,*], Maurice Öhlert[1], Roland Eichinger[2,3], and Christoph Jacobi[1]

[1]Leipzig Institute for Meteorology, Faculty of Physics and Earth Sciences, Leipzig University, Germany
[2]Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany
[3]Charles University Prague, Faculty of Mathematics and Physics, Department of Atmospheric Physics, Prague, Czech Republic
[*]now at: University of Natural Resources and Life Sciences (BOKU), Institute of Meteorology and Climatology, Vienna, Austria

Correspondence: A. Kuchar (ales.kuchar@boku.ac.at)

Abstract. The stratospheric polar vortex (SPV) ~~is a phenomenon comprising~~ comprises strong westerly winds during winter in ~~both hemispheres~~each hemisphere. Especially in the Northern Hemisphere (NH) the SPV is highly variable and is frequently disrupted by sudden stratospheric warmings (SSWs). SPV dynamics are relevant because of both ozone chemistry and its impact on tropospheric dynamics. In this study, we evaluate the capability of climate models to simulate the NH SPV by comparing large ensembles of historical simulations to the ERA5 reanalysis data. For this, we analyze geometric-based diagnostics at 3 pressure levels that describe SPV morphology. ~~Moreover, we assess the ability of the models to simulate SSWs subdivided into SPV split and displacement events.~~ A rank histogram analysis reveals that no model exactly reproduces ERA5 in all diagnostics at all levels. Concerning SPV aspect ratio and centroid latitude, most models are biased to some extent, but the strongest deviations can be found for the kurtosis. Some models underestimate the variability of the SPV area. Assessing the reliability of the ensembles in distinguishing SSWs subdivided into SPV displacement and split events, we find large differences between the model ensembles. In general, SPV displacements are represented better than splits in the simulation ensembles, and high-top models and models with finer vertical resolution perform better. A good performance in representing the geometric-based diagnostics in rank histograms ~~is found to be not necessarily connected to~~ does not necessarily imply reliability and therefore a good performance in simulating displacements and splits. ~~Understanding the biases and improving the~~ Assessing the model biases and their representation of SPV dynamics ~~in climate model simulations can help~~ is needed to improve credibility of climate ~~projections, in particular with focus on polar stratospheric dynamics and ozone~~model projections, for example by giving stronger weightings to better performing models.

1 Introduction

In winter the dynamics of the mid-latitude and polar stratosphere are dominated by the stratospheric polar vortex (SPV). The SPV is a circumpolar band of usually strong westerly winds~~. It forms~~ , forming in autumn due to the cooling of the polar stratosphere. When the stratosphere warms again in spring, the temperature gradient reverses and easterly winds prevail during summer (Holton, 1980). The SPV affects the concentration of ozone over the poles: strong winds are accompanied by lower than average temperatures, ~~that allow~~ allowing the formation of polar stratospheric clouds, where ozone depleting substances are activated (Langematz et al., 2014; Lawrence et al., 2020). Via stratosphere-troposphere coupling (Baldwin and Dunkerton, 2001), the SPV can influence tropospheric circulation patterns, temperatures, and precipitation (Thompson et al., 2002; Butler et al., 2017; King et al., 2019). ~~Due to the impact of the stratospheric circulation on the surface weather~~Hence, uncertainties associated with ~~SPV relates to tropospheric uncertainties in~~ the representation of SPV in models relate to uncertainties in tropospheric climate projections, in particular ~~the windiness over Europe , winter precipitation in~~ with the position of the jet and the precipitation patterns over Europe and the Mediterranean region (Scaife et al., 2012; Zappa and Shepherd, 2017) ~~and~~ as well as with sea level pressure over the Arctic (Simpson et al., 2018). Especially in the Northern Hemisphere (NH), where the SPV is highly variable (Baldwin et al., 2021), the strongest changes happen during so-called sudden stratospheric warmings (SSWs). SSWs are abrupt warmings of the stratosphere ~~, connected with a zonal wind reduction or even reversal during~~ connected with zonal westerly wind reductions. During a so-called major SSW, the wind even reverses to easterly. In the NH, SSWs occur on average about 6 times per decade (Charlton and Polvani, 2007). They are much less frequent in the Southern Hemisphere, with only one recorded major SSW in 2002 since the beginning of the satellite era (Jucker et al., 2021). ~~To distinguish a SSW from the final warming when the wind direction changes to East for the entire summer, an SSW warming must be followed by at least 10 consecutive days of westerly winds (?).~~SSWs can be categorized into two different kinds. Either the SPV is split into two separate SPVs or it is displaced to lower latitudes (Charlton and Polvani, 2007). It is still a matter of current research whether the pressure patterns before the event, and especially the surface pressure response after the event are different depending on ~~the type of SSW~~ SSW type (Mitchell et al., 2013; Seviour et al., 2013; Maycock and Hitchcock, 2015). ~~Particularly for exceptionally strong SPV conditions, dynamical downward coupling from the stratosphere to the troposphere can be observed even influencing surface weather and climate patterns (Black, 2002; Scaife et al., 2005; ?; Baldwin and Dunkerton, 2001). The skill of weather forecasts in the extratropical troposphere is enhanced following an SSW (Sigmond et al., 2013; Tripathi et al., 2015).~~

Multiple studies have focused on analyzing whether climate change alters stratospheric dynamics (Manzini et al., 2014; Ayarzagüena et al., 2020; Rao and Garfinkel, 2021). In the model simulations of the Climate Model Intercomparison Project 5 (CMIP5), the largest uncertainty between individual models regarding a change of stratospheric wind speeds is found at 60°N and 10 hPa (Manzini et al., 2014), the region where the ~~strength of the SPV~~ SPV strength and SSWs are commonly diagnosed (Charlton and Polvani, 2007). In line with this uncertainty, there is no agreement among CMIP5 and CMIP6 models on a ~~possible change in frequency of SSW events~~ trend in SSW frequency (Ayarzagüena et al., 2018; Rao and Garfinkel, 2021). The multi-model mean suggests a slight ~~increase in SSW frequency~~ SSW frequency increase, but the inter-model spread is large, even in the historical simulations (Rao and Garfinkel, 2021). Ayarzagüena et al. (2018) used 12 Chemistry Climate Model Initiative (CCMI) models for their analysis and found that most of them do not project a significant ~~change in SSW frequency~~ SSW frequency trend. Seviour et al. (2016) used two-dimensional ~~diagnostics of the SPV~~ SPV diagnostics to differentiate between vortex splits and displacements and found that most CMIP5 models show some bias in simulating SSWs. Hall et al. (2021) made similar findings with CMIP6 models and found no notable improvement compared to CMIP5. ~~They found that~~ Differences in the chemistry schemes ~~used in the modelsmight influence the results. Other reasons for the large inter-model spread are assumed to be the different~~ of the models, the mean SPV strengths ~~and differences in upward-propagating~~ as well as in upward propagating wave activity flux (Wu and Reichler, 2020) have been identified as possible reasons for the large spread in SSW frequency and its trends.

The large inter-model spread and the uncertainties in the SPV response to climate change underline the need to investigate the reliability of climate models in simulating the SPV. This includes its form, strength and stability. ~~Previous studies mostly~~ Most previous multi-model studies only used a single run from each climate model. ~~Single-model realizations limit analysis as to whether the differences between the models are caused by different model physics or by~~ However, single-model realizations limit possibilities in attributing model differences to the underlying physics or to natural variability (Blanusa et al., 2023; Deser et al., 2020). ~~Thus, particularly the high variability of the~~ Particularly the highly variable wintertime NH stratosphere requires analysis using ~~a large ensemble size~~large ensemble sizes (Deser et al., 2020). ~~In the following, we aim to assess how reliable~~ Therefore, we here assess the reliability of recent large-ensemble model simulations ~~are~~ in representing the SPV and its spatial variability. For this, after introducing our methods in in Sect. 2 we compare in Sect. 3 geometric SPV diagnostics in large climate model ensembles with ERA5 reanalysis data using rank histograms (Matthewman et al., 2009; Seviour et al., 2013). Furthermore, the reliability of the ensembles in detecting SSWs separated into SPV splits and displacements ~~will be~~ is analyzed in Sect. 4. ~~Due to the large SPV variability and the high SSW frequency in the NH, we limit our analysis to the NH SPV.~~ In Sect. 5 we discuss our results with regard to possible reasons for the detected model differences and we finish the paper with some concluding remarks in Sect. 6. Due to the large SPV variability and the high SSW frequency in the NH, we limit our analysis to the NH SPV.

2 Methods

2.1 Data

For this climate model assessment on SPV strength, form and stability, we use large climate model simulation ensembles. Each ensemble consists of multiple simulation members, which only differ by modified initial conditions, ~~otherwise~~ while the model physics and setups are identical (Deser et al., 2020). ~~The advantage of an ensemble in comparison to a climate model with only one realization is that we can estimate the internal variability of a variable from the spread among the individual ensemble members.~~ In our analysis, we use climate models from the Multi-Model Large Ensemble Archive (MMLEA) provided by the US CLIVAR (Climate and Ocean - Variability, Predictability, and Change) working group on large ensembles (Deser et al., 2020) as well as ensembles from the Coupled Model Intercomparison Project 6 (CMIP6, Eyring et al., 2016). We use the historical simulations of ~~these~~ those ensembles where all CMIP5- and CMIP6-class historical forcings are included. Information about the 11 climate ~~model ensembles we used for our analysis can be found~~ models used in our analysis are provided in table 1. The selection criteria for our model database were firstly, availability of at least 10 ensemble members and secondly, availability of the geopotential height at the pressure levels 10, 50 and 100 hPa to calculate the SPV moment diagnostics (see section 2.2)~~at the pressure levels 10, 50 and~~. GFDL-CM3 is an exception, for this model data was available only at 100 hPa.~~However, not all levels are available in all models.~~

For reference, we compare the ~~ensembles of the historical simulations~~ historical simulation ensembles with ERA5 reanaly-

Table 1. Analyzed climate model ensembles from CMIP5 and CMIP6. Low-top models are above and high-top models are ~~separated by~~ below the horizontal line.

Model	Ensemble members	Horizontal resolution (lat x lon)	Center of uppermost level in the vertical	Number of levels in the vertical	Reference
CanESM2~~[5]~~[*]	50	2.8°x2.8°	1 hPa	35	Kirchmeier-Young et al. (2017)
CanESM5	35	2.8°x2.8°	1 hPa	49	Swart et al. (2019)
CESM2	49	0.9°x1.25°	2.26 hPa	32	Danabasoglu et al. (2020)
CNRM-CM6-1	26	1.4°x1.4°	86.4 km	91	Voldoire et al. (2019)
GFDL-CM3~~[5]~~[*+]	20	2.0°x2.0°	86.4 km	48	Donner et al. (2011)
INM-CM5-0	10	2.0°x1.5°	0.2 hPa	73	Volodin and Gritsun (2018)
IPSL-CM6A-LR	32	2.5°x1.25°	80 km	79	Boucher et al. (2020)
MIROC6[#]	10	1.4°x1.4°	0.004 hPa	81	Tatebe et al. (2019)
MPI-ESM1-2-LR	~~30~~50	1.5°x1.5°	0.01 hPa	47	~~Maher et al. (2019)~~Olonscheck et al. (2023)
MPI-ESM1-2-HR	10	0.4°x0.4°	0.01 hPa	95	Müller et al. (2018)
UKESM1-0-LL	16	1.25°x1.9°	85 km	85	Sellar et al. (2019)

[*] Model is part of CMIP5, other models are part of CMIP6.
Only UKESM1-0-LL and CNRM-CM6-1 include full interactive chemistry, CNRM-CM6-1 has a simplified interactive chemistry scheme, the other models are run in dynamics-only mode.
[+] Model data is only available at 100 hPa.
[#] Model is available for 10 ensemble members only. All 50 ensemble members only cover monthly mean data (Shiogama et al., 2023).

sis data (Hersbach et al., 2020)~~. In~~ , or in other words, ERA5 data serves as ground truth in our analysis. ~~We use~~ In this regard, we emphasize that ERA5 is designated as a state-of-the-art benchmark regarding its extensive horizontal and vertical resolution compared to other reanalyses (Hoffmann and Spang, 2022), however, clearly not as an absolute truth. We apply the same geopotential height-based SPV diagnostics ~~for the~~ to the ERA5 reanalysis that we also apply to the model ensemble data. The analysis is carried out for the period 1979–2014 covering years when as many observations as possible were assimilated ~~in~~ to ERA5 including satellite observations (Hersbach et al., 2020). Vokhmyanin et al. (2023) found 22 SSWs in these 36 years of ERA5 data. We analyze daily data of the months from November through March in the NH, as this is when the SPV is usually present and SSW disruptions happen.

2.2 Polar vortex moment diagnostics

To assess spatio-temporal SPV characteristics, the following 2-dimensional moment diagnostics are calculated: aspect ratio, kurtosis, centroid longitude and latitude, and objective area (for details of the calculation see Matthewman et al., 2009; Seviour et al., 2013). The aspect ratio, i.e. the ratio of the major to the minor axis of the SPV ellipse, diagnoses how stretched the SPV is. High/low aspect ratio values indicate a stretched/circular SPV, and exceptionally high values are often associated with

SPV disturbances such as SSWs. The excess kurtosis is a measure for the distribution of geopotential height values inside the SPV, constant geopotential height values lead to an excess kurtosis of 0. A low geopotential height center, i.e. a stable SPV, is represented by high kurtosis values and two separated areas of low geopotential height, i.e. a vortex split, is indicated by a negative kurtosis. The SPV location is diagnosed by centroid latitude and longitude. A lower centroid latitude is often associated with a disrupted SPV and can indicate a displacement, centroid longitude values can additionally help determining the SPV position. The objective SPV area is an indicator for SPV strength, as a large area of low geopotential height is often connected with high wind speeds. More detailed descriptions are provided in Section S2 in the Supplement. In contrast to Matthewman et al. (2009) who described their method using potential vorticity, we ~~will~~ here use the geopotential height to define the SPV edge, as suggested by Seviour et al. (2013). ~~These metrics indicate whether the SPV is stretched, filamented as well as longitudinally and latitudinally displaced, and describe its area, respectively.~~ For this, the algorithms from Seviour et al. (2013) have been modified accordingly, the updated versions can be accessed from Kuchar and Öhlert (2024).

We show the geopotential height climatology at 10 hPa (*gh10*) of all analyzed model ensembles and ERA5 for the period 1979–2014 in Fig. 1. The figure shows that some models do not simulate *gh10* well in comparison to ERA5 (CanESMs, CESM2). On the other hand, visual resemblance between ERA5 and other models (e.g. UKESM1-0-LL) can clearly be seen, too. However, details about the reliability of these large-ensemble model simulations in representing the SPV and its spatial variability cannot be decomposed based on such depictions. Hence, moment diagnostic analyses are needed to shed light on the SPV and its properties in large ensemble simulations.

2.3 Rank histograms

~~We create rank histograms consisting~~ The rank histogram (RH) is a tool used in ensemble forecast verification to determine the reliability of ensemble forecasts and to diagnose errors in its mean and spread. RHs consist of $n+1$ bins, where n is the number of model ensemble members. For ~~this, the values of the ensemble of a certain~~ each time step and variable, the ensemble values are sorted in ascending order~~. For reference,~~ , and the ERA5 reanalysis value at ~~this~~ that particular time step is placed into this set at position k. The histogram counts of all bins greater ~~or equal~~ than or equal to k are then increased by one~~. This~~ , and this procedure is repeated for ~~all available time steps~~ each time step (for details of the calculation see Hamill, 2001; Wilks, 2011). For a reliable (calibrated) ensemble, the counts should be uniformly distributed over all bins. If the ensemble deviates from the reanalysis, the shape of the histogram can be used to find out why (Wilks, 2011). For example, if the historical simulations are biased, there will be a linear trend in the histogram. When the counts of the bins are higher on the left and lower on the right of the histogram, the ensemble simulates the variable to be higher disproportionately often, which is called overforecasting bias. The opposite would be an underforecasting bias. If the ensemble under-/overestimates the variability, the ranks at the ~~sides~~ edges of the histogram have higher/lower counts than in the ~~middle~~center, which is called under-/overdispersion.

For objective assessment of the results, ~~various measures can be considered, such as~~ we consider an additional diagnostic in our study, namely the χ^2 statistic. This diagnostic quantifies how close the ~~rank histogram~~ RH is to an ideal uniform distribution. A perfectly flat histogram would produce a χ^2 value of 0. Jolliffe and Primo (2008) introduced a method to split the χ^2 statistic into multiple metrics, where each one describes a ~~particular shape of the histogram~~ certain histogram shape. The linear

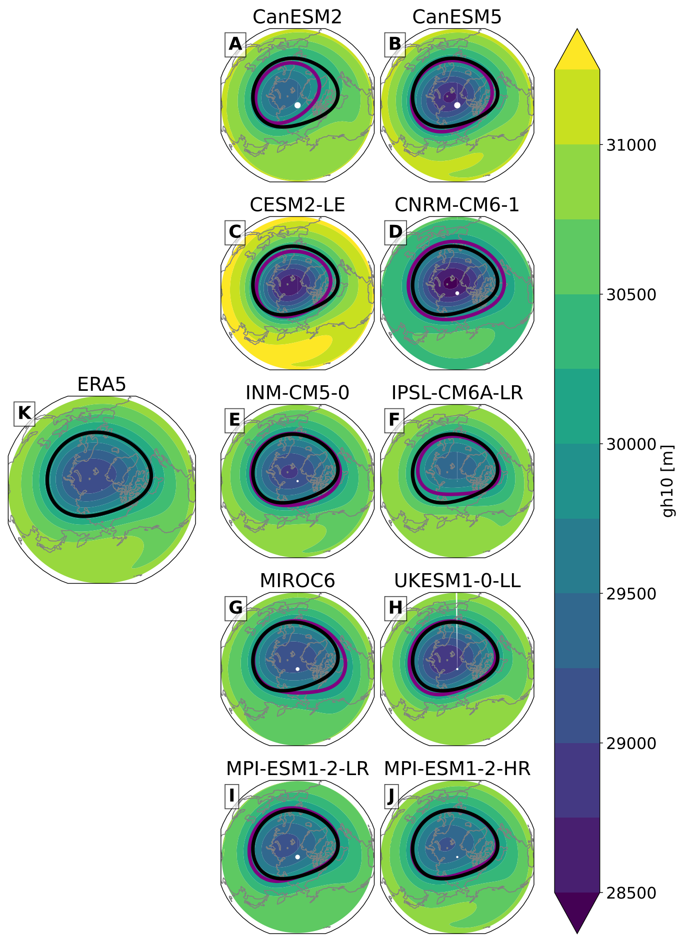

Figure 1. Geopotential height climatology at 10 hPa (*gh10*) of all analyzed model ensembles: CanESM2 (**A**), CanESM5 (**B**), CESM2 (**C**), CNRM-CM6-1 (**D**), INM-CM5-0 (**E**), IPSL-CM6A-LR (**F**), MIROC6 (**G**), UKESM1-0-LL (**H**), MPI-ESM1-2-HR (**I**) and MPI-ESM1-2-LR (**J**), and ERA5 (**K**) for the period 1979–2014. The black and purple line represents contour of 30000 m in ERA5 and in a particular model, respectively.

trend ~~can be~~ is used as a bias indicator and ~~the~~ a U-shape indicates over- or underdispersion (spread). These metrics can be especially helpful when both bias and over- or underdispersion are present in an ensemble, as this can be difficult to distinguish ~~from the rank histogram alone.~~ visually from the RH alone. The contributions of these two components to the total χ^2 statistic are presented along all RHs in our assessment. These statistics should serve in relation to the other models instead of defining any threshold for a "good" or "bad" model.

2.4 Perfect model range

Due to internal variability, it is possible that a ~~rank histogram~~ RH has a somewhat uneven distribution~~, even though the ensemble is covering the probabilities correctly. This can even result in an insignificant χ^2~~. To determine which deviations from a uniform distribution can be attributed to internal variability, Suarez-Gutierrez et al. (2021) suggested the use of 'a perfect model range'. To obtain this range, a rank histogram is created for each ensemble member where this specific member is treated as a reference (i.e., as if it was the reanalysis). This results in slightly different values for each bin in the ~~rank histograms~~RHs, depending on the member in question. The perfect model rank range is then defined by the range where 90% (5th - 95th percentile) of the bin counts are found. Since a member from the ensemble can never be higher or lower than all ensemble members, the values for the rank range in the first and last bin are ignored.

2.5 SSW diagnostics

~~SSWs~~ SSW events can be subdivided into SPV splits and displacements and can be detected by means of the metrics described above. ~~Seviour et al. (2013) suggested that an SPV split can be detected by using the aspect ratio and they defined values~~ As suggested by Seviour et al. (2013), we detect SPV splits by an aspect ratio higher than 2.4~~to be a split~~. For a displacement, the centroid latitude is arguably the best indicator. Here, ~~Seviour et al. (2013) defined that a displacement has taken place~~ as defined by Seviour et al. (2013) a displacement is detected if the centroid latitude is lower than 66^{o}N.

To assess how well the probability of these events is represented in the model simulation ensembles, the receiver operating characteristics (ROC) curves are used (see Figs. S1 and S2 and their description in Section S3 in the Supplement). The area under the ROC curve (AUC) indicates how well an ensemble is able to discriminate between SSW and non-SSW events using the thresholds above (with reference to ERA5)~~: It~~. AUC ranges from 0 to 1. ~~A value~~ Values of 1~~and~~, 0.5 ~~indicates perfect skill~~and 0 indicate perfect skill, random guessing, ~~respectively.~~ and no skill, respectively. As an example, we show ROC for displacements and splits in Fig. 2 for the CanESM5 ensemble. Bin values indicated along ROC are probabilities of whether the model simulates displacements and splits across its ensemble members, respectively. These values then serve as inputs for the calculation of dichotomous contingency tables which includes true and false positive rates displayed on y- and x-axis, respectively. Fig. 2**B** demonstrates that CanESM5 cannot discriminate split events better than random guessing (see Sect. 4.2). To determine the uncertainty of the AUC, we provide error bars using the approach of the perfect model range, where we assume each ensemble member ~~as~~ an observation.

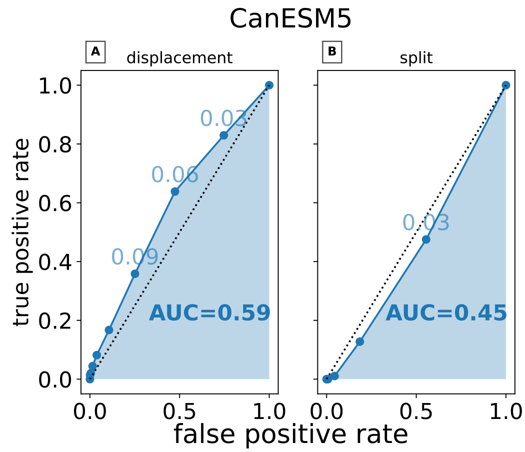

Figure 2. Receiver operating characteristics (ROC) curves for displacements (**A**) and splits (**B**) in CanESM5. Bin values indicated along ROC are probabilities whether model simulates displacements and splits across its ensemble members, respectively. Area under the ROC curve (AUC) is visualized and also specified in the figure. The dashed line represents random discrimination skill, i.e. AUC=0.5.

We also reproduce the methodology from Hall et al. (2021) as previously applied in Mitchell et al. (2011) and based on Seviour et al. (2013) to examine relationships between modal centroid latitude and aspect ratio and displacement and split SSW frequency, respectively. The frequencies of ERA5 SSW split and displacement events determined with this method are within the uncertainty of other methods (∼6.94 events per decade, ∼6.66 events per decade including displacement and splits events only, displac . We provide the list of ERA5 SSW split and displacement events in Tab. S2 in the Supplement to document this agreement.

3 Analysis of geometric polar vortex diagnostics

In the following, the agreement between the ERA5 reanalysis and the historical simulations of the climate model ensembles will be compared by means of the SPV moment diagnostics ~~that were introduced in section 2.2. Rank histograms~~ introduced in Sect. 2.2. For this, RHs are discussed for all available models at 10, 50 and 100 hPa, however, the figures for 50 hPa and for 100 hPa are shown in the Supplement. We also summarize bias and spread from these figures in Table S1 in the Supplement.

3.1 Aspect ratio

The aspect ratio ~~is~~ of the SPV is determined by the ratio of the major to the minor ellipse axis. Thus, it measures how stretched the SPV is providing high values for stretched and low values for more circular SPVs. Exceptionally high aspect ratio values are of particular interest for SPV dynamics ~~because it is an indicator for vortex~~ as they are often associated with SPV disturbances, in particular indicating splitting events (Seviour et al., 2013). Fig. 3 shows ~~rank histograms~~ RHs of the aspect ratio for all analyzed climate model ensembles together with the above ~~explained~~ introduced statistical values χ^2, bias and spread at

10 hPa, the level that is most commonly used to detect SPV splits. As indicated in Sect. 2, interpretation of all results here are with reference to ERA5 reanalysis data~~, which is considered as ground truth for the analysis.~~ .

All models succeed to simulate the spread of the aspect ratio, but most models are biased to some extent. At 10 hPa, four models are biased ~~simulating~~ and simulate lower aspect ratios more frequently than the reanalysis (CanESM2, CanESM5, CESM2 and CNRM-CM6-1) and ~~four~~ three models show a ~~high~~ overforecasting bias (IPSL-CM6A-LR ~~, MIROC6,~~ and both MPI-ESM1-2 ensembles). Only ~~two~~ three models show no ~~significant~~ considerable bias (INM-CM5-0, MIROC6 and UKESM1-0-LL).

The strongest biases are found in the ~~CanESM2~~ CanESM5 and CESM2 ensembles ~~. Such strong biases can have implications for the simulation of SPV splits. Frequent simulation of low aspect ratios points towards an~~ (see also Tab. S1 in the Supplement). These relatively strong aspect ratio biases (as compared to the other models) point towards underestimation of SPV split probability, ~~which will further be investigated in~~ but see Section 4.2 ~~.~~ for further investigation on this connection.

~~The~~ With the rank histograms at 50 and 100 hPa ~~can be found in the supplement (Figs. S1, S6). The models generally~~ (see supplementary Figs. S3, S8), the models can be separated into ~~three~~ two groups according to their behavior ~~compared to~~ in relation to the results at 10 hPa. One group of models shows larger biases at lower altitudes (CanESM2, INM-CM5-0 and UKESM1-0-LL). In ~~INM-CM5-0 and UKESM1-0-LL the low bias in aspect ratios appears only at the two lower levels. In the other~~ the other model ensembles the bias is weaker at lower altitudes (CanESM5, CESM2, INM-CM5-0, IPSL-CM6A-LR, MIROC6 and MPI-ESMs).

3.2 Centroid latitude

Fig. 4 shows the ~~rank histograms~~ RHs of the centroid latitude for all analyzed climate model ensembles at 10 hPa. The centroid latitude is a measure of how far the polar vortex is shifted from the North Pole; untypically low latitudes indicate vortex displacements (Seviour et al., 2013). Most ensembles show a bias in centroid latitude, but the spread is generally represented well. This is similar to the results ~~of~~ for the aspect ratio. The direction of the biases is not consistent among the models. The CanESM2, IPSL-CM6A-LR and MPI-ESM1-2 ensembles simulate a ~~high~~ low latitude bias with regard to the reanalysis, while the CESM2, INM-CM5-0 and MIROC6 ensembles show a ~~low~~ high latitude bias. Only CanESM5, CNRM-CM6-1 and UKESM1-0-LL do not show any notable bias or spread, i.e., the corresponding statistical diagnostics show low values. Here, CanESM5 (see Fig. 4**B**) shows a notable improvement compared to its earlier version CanESM2 (see Fig. 4**A**). The combination of ~~biases towards higher centroid latitudes and lower aspect ratios~~ high centroid latitude bias and low aspect ratio can only be seen in CESM2 (see Fig. 4**C**), which can explain the general underestimation of SSWs in this model (see Sect. 4 and 5).

The ~~rank histograms~~ RHs of the centroid latitude at the two lower ~~pressure~~ altitude levels are shown in Figs. ~~S2 and S7~~S4 and S9 in the Supplement. In most models, the bias is similar at ~~the different levels , which means~~ all analyzed levels (10, 50 and 100 hPa), showing that the performance ~~of most ensembles~~ with respect to centroid latitude is not very sensitive to altitude. Only in CanESM5~~a bias towards lower latitudes~~ , a low latitudes bias appears at the two lower ~~levels~~altitudes. In MIROC6 the ~~bias to simulate lower latitudes~~ low latitude bias is only present at 10 and 50 hPa~~;~~ , while at 100 hPa the ~~bias vanishes completely and the histogram is almost perfectly flat~~ model shows almost no bias and thus a nearly perfectly flat histogram.

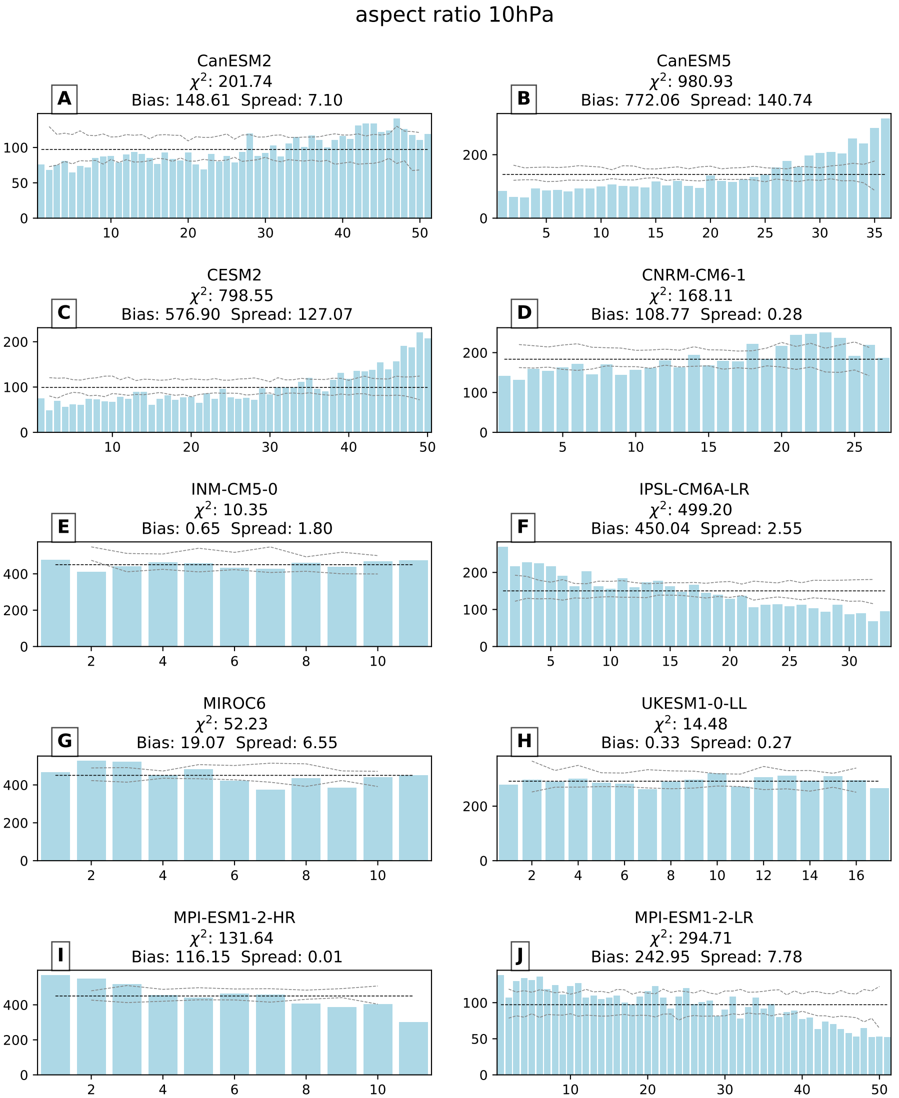

Figure 3. Rank histograms of aspect ratio at 10 hPa of all analyzed ~~climate~~ model ensembles~~with their respective statistics~~: CanESM2 (~~see Sect. 2~~**A**), CanESM5 (**B**), CESM2 (**C**), CNRM-CM6-1 (**D**), INM-CM5-0 (**E**), IPSL-CM6A-LR (**F**), MIROC6 (**G**), UKESM1-0-LL (**H**), MPI-ESM1-2-HR (**I**) and MPI-ESM1-2-LR (**J**). Blue bars show counts for the individual bins, the black dashed line corresponds to the expected value for a flat histogram, gray dashed lines indicate the perfect model range (see Sect. 2.4). The x-axis shows the ensemble member number and the y-axis shows the count of the bins. The contributions of bias and spread to the total χ^2 statistic are provided above the rank histograms for each model (see Sect. 2).

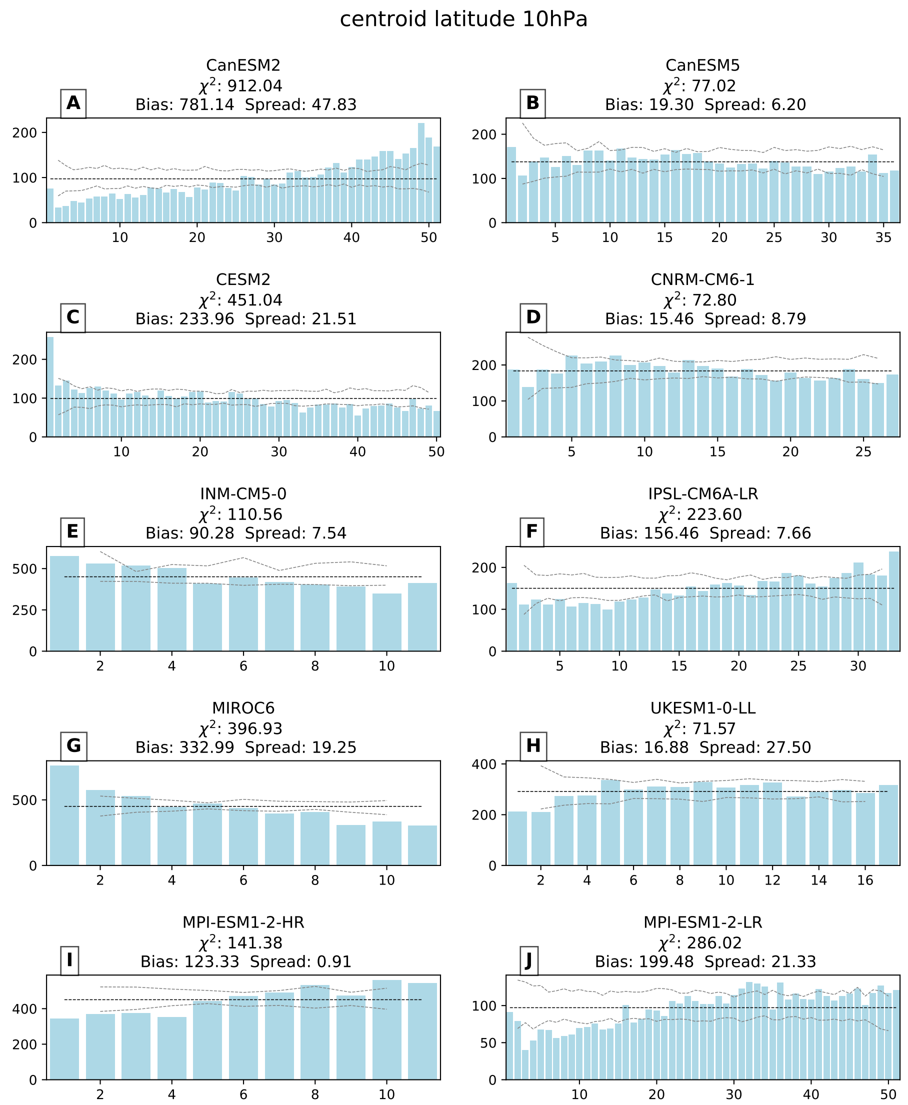

Figure 4. As Fig. 3, but for centroid latitude at 10 hPa of all analyzed model ensembles: CanESM2 (**A**), CanESM5 (**B**), CESM2 (**C**), CNRM-CM6-1 (**D**), INM-CM5-0 (**E**), IPSL-CM6A-LR (**F**), MIROC6 (**G**), UKESM1-0-LL (**H**), MPI-ESM1-2-HR (**I**) and MPI-ESM1-2-LR (**J**).

3.3 Centroid longitude

The centroid longitude in the climate models ranges from -180° to +180°, the negative values lie in the Western Hemisphere and the positive ones in the Eastern Hemisphere. The centroid longitude ~~rank histograms~~ RHs (Fig. 5) show where the climate models over- or underestimate the position of the SPV. When the counts are lower/greater than average, the ensemble simulates the SPV center more/less frequently at the respective longitude. The centroid longitude is depicted best by the CNRM-CM6-1 (see Fig. 5**D**) and INM-CM5-0 (see Fig. 5**E**) ensembles. ~~Other~~ The other ensembles show notable deviations from a flat histogram, but ~~there is no consistency in the deviations among the different~~ these deviations are not consistent among the models. The CanESM2, CESM2 and UKESM1-0-LL ensembles simulate the SPV center in the Eastern Hemisphere more frequently than the reanalysis. The IPSL-CM6A-LR and MIROC6 ensembles show the opposite bias. The ~~rank histograms~~ RHs of CanESM5 (see Fig. 5**B**) and the MPI-ESMs (see Fig. 5**I-J**) show lower counts on both ends, indicating that the ensembles simulate the SPV more frequently in and around the region of the Bering Strait and Alaska (the meridian of +180/-180°) than the reanalysis. ~~The rank histograms of the centroid longitude at~~

At 50 and 100 hPa ~~are shown in~~ (Figs. ~~S3 and S8. Some models show a change in the direction of bias or~~ S5 and S10 in the Supplement), some models show biases of opposite sign than at 10 hPa, or even a general dispersion at the 3 different pressure levels (e.g. CanESM2, IPSL-CM6A-LR, MIROC6). This could not be seen for the centroid latitude. In most models both bias and spread are present at least ~~in~~ at some pressure levels. Only CNRM-CM6-1 produces a flat histogram where almost all counts lie inside the perfect model range in all ~~three~~ 3 pressure levels. The UKESM1-0-LL and MPI-ESM1-2 ensembles show flat histograms at 50 and 100 hPa.

3.4 Kurtosis

The excess kurtosis is a measure for how the values of geopotential height are distributed within the SPV region (Matthewman et al., 2009). Mitchell et al. (2011) proposed that this diagnostic can be used to detect both SPV split and displacement events. They showed that exceptionally low values are often ~~a sign~~ an indication that an SPV split has occurred~~, high~~. High positive values on the other hand can occur after splits and displacements (see their Figs. 2 and 5). The ~~rank histograms~~ RHs for the excess kurtosis at 10 hPa are shown in Fig. 6. Four models show similar ~~rank histograms~~ RHs with much higher counts on the left side of the histogram, namely CanESM2, CanESM5, CESM2 and CNRM-CM6-1. These ensembles underestimate the variability of the kurtosis and additionally simulate a kurtosis ~~high~~ positive bias. The result is that very low values of the kurtosis are simulated much less frequently in the models than they occur in the reanalysis. Therefore, ~~these models likely underestimate~~ this may contribute to the underestimation of the SPV split frequency. This is in line with the aspect ratio low bias in these models (see Sect. 3.1), except for CNRM-CM6-1 (see Fig. 6**D**). The UKESM1-0-LL (see Fig. 6**H**) ensemble shows a similar kurtosis behavior as the four model ensembles ~~from~~ mentioned above, but the deviations are by far not as pronounced. Even though the UKESM1-0-LL ensemble shows a good representation of the aspect ratio, it simulates low values of the kurtosis less frequently. This is in line with the results by (Hall et al., 2021), who reported that the model simulates too few split events. The MIROC6 and INM-CM5-0 ensembles perform best in representing the kurtosis, in particular at 10 hPa. ~~Despite~~

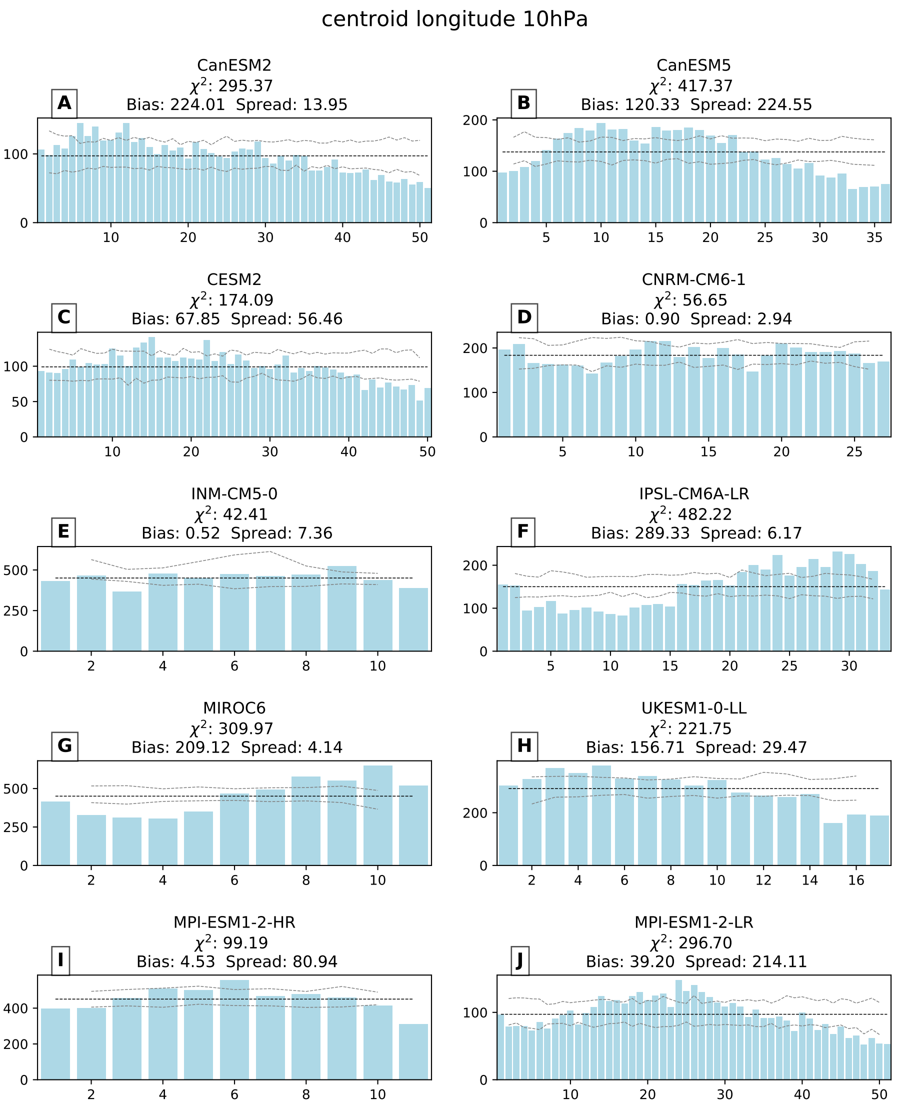

Figure 5. As Fig. 3, but for centroid longitude at 10 hPa of all analyzed model ensembles: CanESM2 (**A**), CanESM5 (**B**), CESM2 (**C**), CNRM-CM6-1 (**D**), INM-CM5-0 (**E**), IPSL-CM6A-LR (**F**), MIROC6 (**G**), UKESM1-0-LL (**H**), MPI-ESM1-2-HR (**I**) and MPI-ESM1-2-LR (**J**).

the differences in vertical and horizontal resolution, there is no substantial difference between the two Both MPI-ESM1-2 simulations . However, we conclude that if the spatial model resolution has a substantial influence on the representation of the SPV, dedicated experiments with systematic resolution changes will be needed in future studies. contains bias but a dome-shaped RH in MPI-ESM1-2-LR indicates large ensemble spread.

At the two lower altitudes (Figs. S4 and S9S6 and S11 in the Supplement), most ensembles underestimate the kurtosis variability (CanESM2, CanESM5, CESM2, INM-CM5-0 and UKESM1-0-LL and GFDL-CM3 at 100 hPa). Only the IPSL-CM6A-LR and both MPI-ESM1-2 ensembles overestimate it. The MPI-ESM1-2 ensembles show almost flat rank histograms RHs at 100 hPa. Generally, in comparison with centroid latitude or aspect ratio most models do not most models simulate the kurtosis well with respect to ERA-5less well than centroid latitude or aspect ratio, in particular at 10 hPa. This suggests that centroid latitude and aspect ratio seem to be are more reliable indicators for SPV split and displacement SSW frequency estimates than the kurtosis is.

3.5 Objective area

The objective area is of interest because a larger/smaller area of low geopotential height is often related to a stronger/weaker SPV with higher/lower wind speeds. Multiple ensembles (INM-CM5-0, IPSL-CM6A-LR, MIROC6 and both high and low resolution MPI-ESM1-2 ensembles) simulate a strong small SPV negative bias at 10 hPa (Fig. 7). In addition to that, these models underestimate the variability of the objective area. This combination results in a strong underestimation of a large SPV occurence. The CanESM2 (see Fig. 7**A**) ensemble also shows this combination, but not as pronounced as the above-mentioned models. The CanESM5, CESM2 , CNRM-CM6-1 and UKESM1-0-LL and CNRM-CM6-1 ensembles simulate a large SPV positive SPV area bias, which is likely connected to too high wind speeds.

At the lower altitudes (see Figs. S5 and S10S7 and S12 in the Supplement) most models are biased in the same direction as at 10 hPa, but in some models the strength of the bias varies with height. Only INM-CM5-0 shows a small SPV area bias at 10 hPa (see Fig. 7**F**) and a large SPV area bias at 50 hPa. At 100 hPa barely any bias can be detected in INM-CM5-0. Models with a large SPV area bias at 100 hPa , also simulate a low aspect ratio bias at 10 hPa (CanESM2, CanESM5, CESM2) and vice versa (IPSL-CM6A-LR and both MPI-ESM1-2 ensembles - except for MIROC6). Models without any notable bias (irrespective of the spread) for the objective area show a good representation of the aspect ratio (CNRM-CM6-1, INM-CM5-0, UKESM1-0-LL). A connection between lower weaker stratospheric winds and SPV split frequency in the CMIP6 models was already noted by Hall et al. (2021). They found that the frequency of SPV splits was related to the wind speeds at 100 hPa, because higher wind speeds hinder the upward propagation of wave number 2 planetary waves into the stratosphere. Similarly, Wu and Reichler (2020) showed that the highest uncertainty in the frequency of SSWs SSW frequency comes from an uncertainty in lower stratospheric wind speeds. Most In general, most models do not succeed to simulate the objective area of the SPV adequatelySPV area well. Most commonly, they underestimate the variability and additionally a biasis often present. The best representation at often simulate a bias. At 10 hPa is shown by the UKESM1-0-LL ensemble . For represents the SPV area best. At the two lower levelsaltitudes, CNRM-CM6-1 is the closest to a perfect representation, confirmed shows the best representation, depicted by the lowest values of the χ^2-statistic.

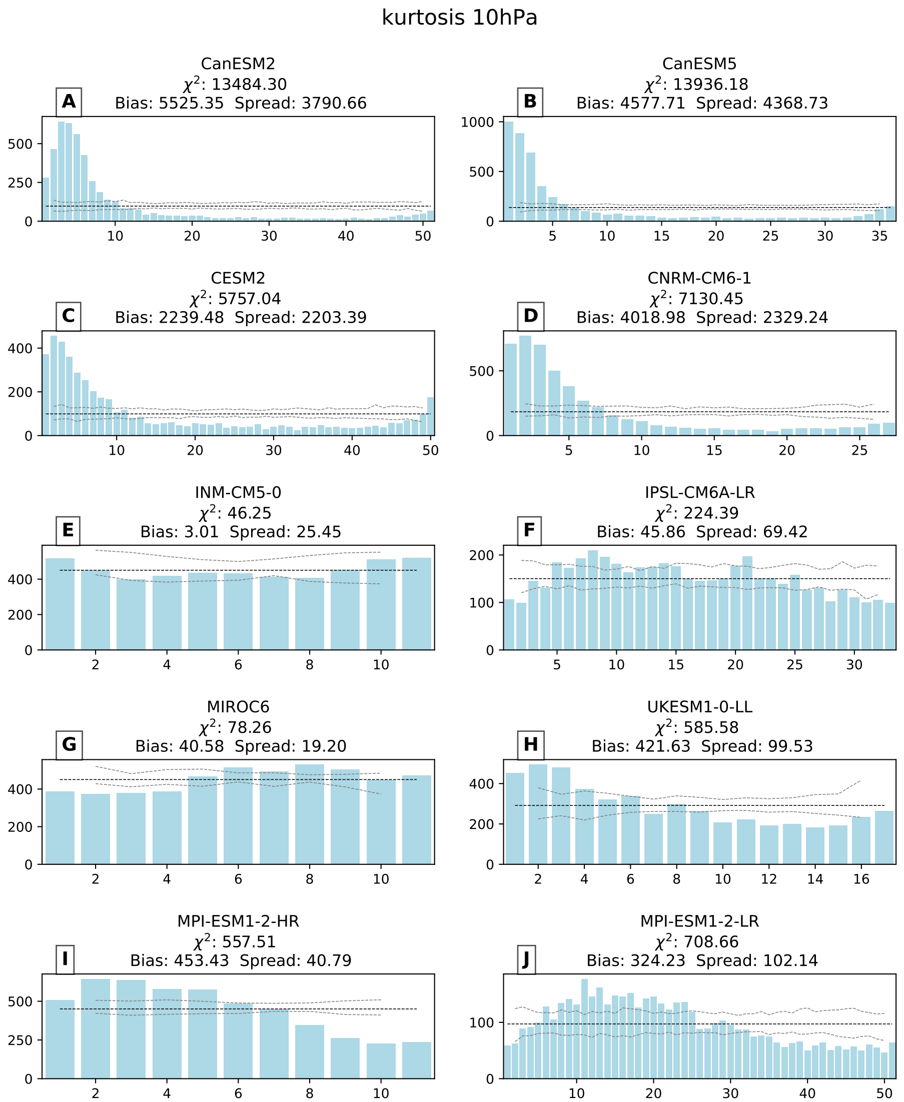

Figure 6. As Fig. 3, but for kurtosis at 10 hPa of all analyzed model ensembles: CanESM2 (**A**), CanESM5 (**B**), CESM2 (**C**), CNRM-CM6-1 (**D**), INM-CM5-0 (**E**), IPSL-CM6A-LR (**F**), MIROC6 (**G**), UKESM1-0-LL (**H**), MPI-ESM1-2-HR (**I**) and MPI-ESM1-2-LR (**J**).

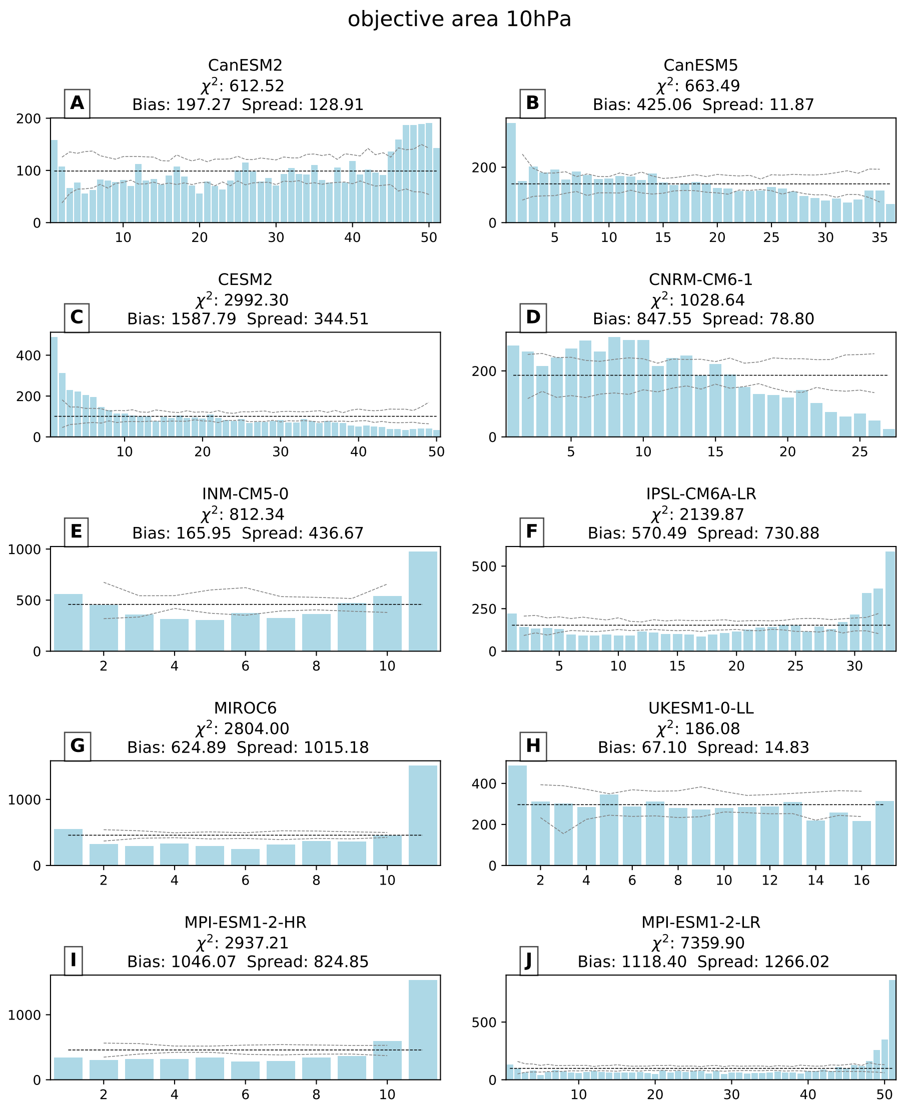

Figure 7. As Figure 3, but for objective area at 10 hPa of all analyzed model ensembles: CanESM2 (**A**), CanESM5 (**B**), CESM2 (**C**), CNRM-CM6-1 (**D**), INM-CM5-0 (**E**), IPSL-CM6A-LR (**F**), MIROC6 (**G**), UKESM1-0-LL (**H**), MPI-ESM1-2-HR (**I**) and MPI-ESM1-2-LR (**J**).

4 Sudden stratospheric warming analysis

~~When studying the NH SPV, a particular focus lies on events when the SPV is disrupted~~A particular focus in NH SPV studies lies on disruptive SPV events, so called sudden stratospheric warmings (SSWs). We here assess the ability of the climate models to distinguish between SSWs and steady SPV conditions. While the ~~rank histograms~~ RHs reveal reliability (consistency), they do not evaluate statistical resolution (the degree to which a forecast sorts the observed events into different groups), so our study needs to be accompanied with other tools such as the ROC (Hamill, 2001). Using the ROC we can analyze how well the different models are able to simulate SSW events through diagnostics of the SPV morphology. The applied method allows us to individually diagnose displacement- and split-type SSWs. Hence, we conduct two separate analyses here, as it has been shown that these events, as well as their surface impact, fundamentally differ (Baldwin et al., 2021, and references therein).

4.1 Displacement events

Fig. 8 shows the areas under the ROC curves (AUC) of the analyzed climate models for detection of SPV ~~splits and displacements. All models reach a value of greater than 0.5 which means their performance is better than randomly guessing. As hypothesized (Hall et al., 2~~ ~~, SPV~~ displacements (see Fig. 8**A**) ~~may be less frequent in models that simulate a high centroid latitude bias.However, the ROC diagram is insensitive to certain types of biases (Kharin and Zwiers, 2003), since a biased model may still have good statistical resolution. However, the ROC can still be considered as a potential skill when the model is correctly calibrated (Wilks, 2011) .This fact can also explain the differences between CanESM2 and CanESM5. According to the AUC, CanESM2 distinguishes between displacement events and stable SPV conditions better than CanESM5. This is remarkable as CanESM5 was performing much better in the rank histogram analysis for the centroid latitude at 10~~and splits (see Fig. ~~hPa compared to CanESM2~~8**B**). For all ROC curves see Figs. S1 and S2 in the Supplement. We also summarize AUC values from these figures in Table S1 in the Supplement.

~~All~~ In general, the low-top models (see Tab. 1) ~~generally~~ reveal lower AUC values than the high-top models, with the exception of MPI-ESM1-2-HR. In fact, MPI-ESM1-2-HR has the lowest value of all analyzed models. Additionally, the AUC for the MPI-ESM1-2-HR ensemble is slightly lower than for its low resolution counterpart ~~:~~ MPI-ESM1-2-LR. The CNRM-CM6-1 ensemble shows the best performance regarding the simulation of SPV displacement events. This model also has one of the best representations of the centroid latitude at 10 hPa in the ~~rank histograms~~RHs. In fact, the ~~rank histogram~~ RH was similar to CanESM5, for which a rather weak performance in the ROC curves was found. A reasonable performance is shown by the INM-CM5-0 ensemble, which in fact has the second highest AUC. IPSL-CM6A-LR ~~, MIROC6 and MPI-ESM1-2-LR~~ and MIROC6 show similar values for the AUC. The UKESM1-0-LL ~~ensemble shows an average~~ and MPI-ESM1-2-LR ensembles show an above-average performance compared to the other climate models. Again, this stands partly in contrast to the rank histograms where ~~this~~ the UKESM1-0-LL ensemble was closest to a flat histogram of centroid latitude with the lowest χ^2 statistic of all models.

These results demonstrate that even if the ~~rank histogram~~ RH implies a good representation, i.e. a reliable ensemble, it can show a comparatively low statistical resolution in distinguishing between displacements and non-displacements (e.g CanESM5)

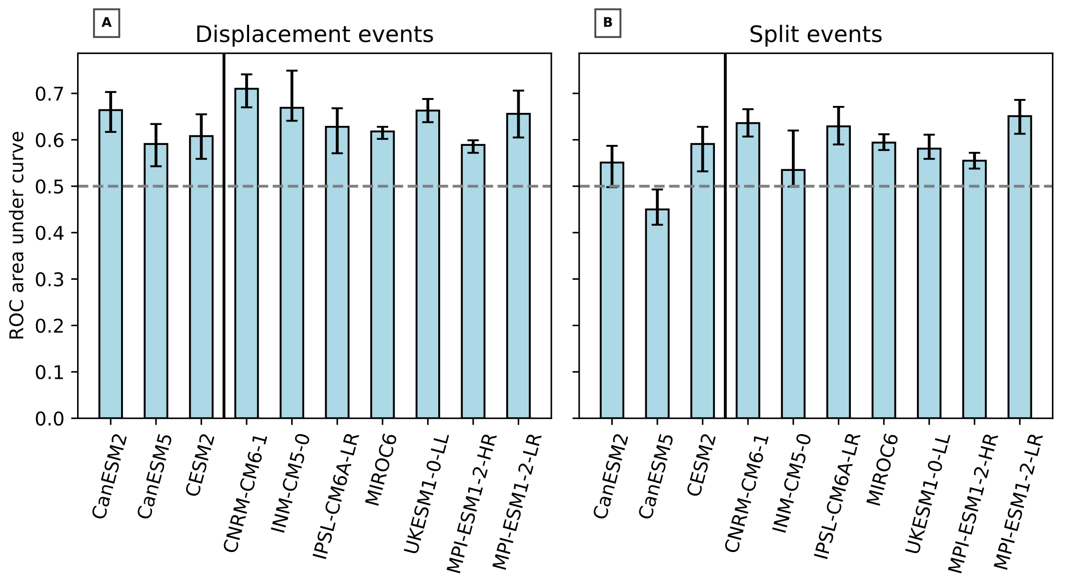

Figure 8. Area under the curve for the ROC curves of the analyzed climate ensembles for displacement (~~left~~**A**) and split (~~right~~textbfB) events. Grey line lies at 0.5, the value at which the simulation is not better than randomly guessing. Low-top models are separated on the left with a black line. Error bars indicate 5th and 95th percentile estimated by using the perfect model range.

and vice versa (e.g. CESM2). Generally, for most of the ensembles the AUC lies in a narrow region around 0.6, implying that the simulation of displacement events can still be improved in the climate ensembles, e.g. by calibration~~.~~ (Wilks, 2011).

It has been suggested by Seviour et al. (2016) and Hall et al. (2021) that models with a bias in centroid latitude also have a bias in displacement frequency in the respective direction. ~~However, Fig. S11 shows that we could not reproduce a clear~~ While we reproduce this negative relationship between number of displacements and modal centroid latitude from Fig. 3a in Hall et al. (2021) ~~. As centroid latitude diagnoses the climatological mean position, and the displacements particular extreme events of this position, it is reasonable that there is no clear relation, although the diagnostics are connected.~~in our Fig. 9**A**, we observe models that despite their biases in modal centroid latitude simulate a comparable frequency of displacement SSWs. The CNRM-CM6-1 ensemble is again the best-performing one in terms of the frequency of displacement SSWs comparable to ERA5 (∼4 events per decade).

4.2 Split events

In general, the climate model ensembles do not simulate SPV splits (see Fig. 8**B**) as well as SPV displacements (see Fig. 8**A**)~~In fact~~. Indeed, all models except for IPSL-CM6A-LR have lower AUCs for split events than for displacements. Overall, a weaker performance of the low-top models can be detected, which is particularly obvious for CanESM5. ~~As in the displacement analysis, CanESM2 performs better than its newer counterpart CanESM5. In the latter the AUC is even lower than~~ The ensemble performs worse than its older counterpart CanESM2 and shows a ROC area of below 0.5~~meaning that in CanESM5~~

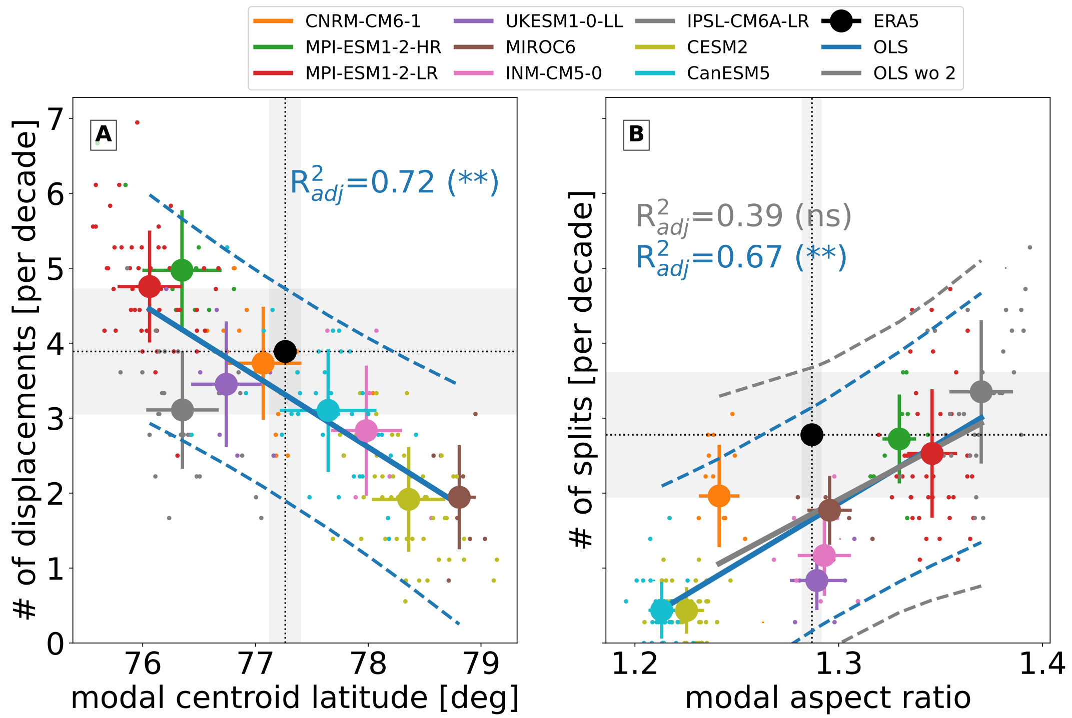

Figure 9. Scatter plots of modal centroid latitude [deg] and frequency of displacement SSWs (**A**), and modal aspect ratio and frequency of split SSWs (**B**) [per decade] in large ensembles compared with ERA5. Blue solid dashed lines are ordinary least square regressions and their 95% confidence intervals for all models, while grey lines in panel (**B**) show regression lines and confidence intervals for high-top models only (i.e. except CESM2 and CanESM2). Dotted lines represent results for ERA5. Horizontal shading indicates the frequency of displacement or split events and represents the 1σ range, assuming a binomial distribution of events. Vertical shading was calculated using bootstrapping of ERA5 time series and represents the 1σ range. The error bars represent standard deviation through ensemble members shown as dots. R^2_{adj} represents adjusted coefficient of determination. The asterisks flag levels of significance with a p-value less than 0.01.

, which means that the false positive rate for detecting split events is higher than the true positive rate. In fact, CanESM5 is the only model that produces an AUC of lower than 0.5 for split events (see also Fig. 2), even when considering the error bars. A reason for this might be the strong bias to lower aspect ratio at 10 hPa that is likely resulting in underestimation of the frequency of SPV splits, which are associated with exceptionally high aspect ratios. The Although the CESM2 ensemble shows a similarly strong bias compared to aspect ratio bias as CanESM5but , it has a better representation of split events according to the ROC plots.

As for displacement events, CNRM-CM6-1 reaches also reaches one of the largest AUC as was already seen for the displacement eventsfor splits after MPI-ESM1-2-LR (see also Tab. S1 in the Supplement). Thus, this model can be regarded to have the best representation of SPV displacements as well as splits and likely SSW events in general (see also Fig. 9). The AUC of the INM-CM5-0 ensemble reaches a value of slightly above 0.5, indicating that the simulation of splits is only marginally better than randomly guessing their occurrence. This result stands in contrast to the fact that this ensemble has shown one of the best performances in the ~~rank histogram~~ RH analysis for the aspect ratio without any significant bias or spread. This again corresponds to the insensitivity of the ROC to certain biases as discussed above. The IPSL-CM6A-LR ensemble on the other hand almost reaches the performance of CNRM-CM6-1 in spite of its bias to predict higher aspect ~~ratio more~~ ratios too often. MIROC6 ~~, MPI-ESM1-2-LR~~ and UKESM1-0-LL show similar AUC that lie between the best and weakest performing ensembles. The high resolution MPI-ESM1-2 ensemble is showing a lower AUC than the low resolution version MPI-ESM1-2-LR, as it was already seen for displacement events, but it still remains well above the value of 0.5.

~~Similar to displacement events,~~ In accordance with the results by Seviour et al. (2016) and Hall et al. (2021)~~found that~~ , we tried to reproduce whether models showing a strong bias to lower aspect ratios in our analysis indeed underestimate the SPV split frequency. However, results in Fig. ~~S12~~ 9**B** are more dispersed compared to Hall et al. (2021). It reveals that the linear relationship between number of splits and modal aspect ratio from Fig. 3b in Hall et al. (2021) cannot be reproduced in the large ensemble simulations of the here used high-top models~~, only~~ . It only works to some degree when low-top models are included. ~~This underlines our results~~ Unlike their results, the reanalysis values lie within the 95% confidence interval of the ordinary least squares fit. As we can rule out that the size of ensemble members might not be sufficiently large in our study, we argue that the fit may not be so robust for split SSWs because a stretching tendency of polar vortex is accompanied with a centroid-latitude tendency to equatorward values (Mitchell et al., 2011). This finding also underlines our statement that good performance in representing the geometric-based diagnostics in ~~rank histograms~~ RHs is not necessarily connected ~~to~~ with a good performance in simulating displacements and splits. Wu and Reichler (2020) also demonstrated that bias-corrected models for vortex strength may not consistently align with reanalyses in terms of revealing SSW frequency.

5 ~~Discussion~~ Summary and ~~Outlook~~discussion

We assessed the SPV in large CMIP5 and CMIP6 climate model ensembles using ~~rank histograms~~ RHs with reference to ERA5 reanalysis data. The performance of the models varies depending on the analyzed variables and pressure levels. No model ensemble can be highlighted as having the best or worst performance over all variables and pressure levels. If the general performance over all levels and variables is regarded, the CNRM-CM6-1 and UKESM1-0-LL ensembles can be considered to be representing SPV form and variability best. These two models produce a flat ~~rank histogram~~ RH for most of the geometric variables at most altitudes, which means that the simulated SPV in these models agrees well with that of ERA5. The flat ~~rank histrogram~~ RH is a necessary but not a sufficient condition for concluding reliability in SPV simulation.

Furthermore, we used the ROC analysis in order to assess the ability of the ensembles regarding SPV displacement and split frequencies. As all models reach an area under the ROC curve of more than 0.5 (see Fig. 8**A**), they distinguish between SPV displacements and non-displacements better than random guessing. ~~The~~ In general, the ensembles represent displacement events better than split events~~in general~~. The best representation of both SPV splits and displacements has CNRM-CM6-1. This model performs well in the ~~rank histogram~~ RH analysis as well. However, a general rule of thumb that connects the ~~rank histogram~~ RH with the ROC analyses could not be found here. This is due to the insensitivity of the ROC to biases in the forecast. The ROC diagram can be considered as a measure of potential usefulness when a model ensemble is correctly calibrated (Wilks, 2011). This can lead to a more reliable forecast while maintaining good discrimination. A joint analysis of variety diagnostics provides the bigger picture about the quality of large-ensemble model simulations.

The model top height seems to be a key factor. Low-top models reveal strong biases for most variables, in particular CESM2. ~~Low-top models were already found to simulate fewer SSWs than in the reanalysis and a lower~~ Charlton-Perez et al. (2013) and Hall et al. (2021) already found that low-top models simulate too few SSWs and a too low variability of the SPV wind speeds ~~(Charlton-Perez et al., 2013; Hall et al., 2021). A better vertical resolution also improves the simulation of the SSW frequency (Wu and Reichler, 2020). The models with the finest spatial resolution (INM-CM5-0 and MIROC6) showed a good performance, especially for the aspect ratio and kurtosis at 10 hPa~~. Overall, this means that the downward influence of the upper stratosphere and mesosphere has a large influence on the SPV and on SSWs ~~(Hitchcock and Simpson, 2014)~~ (see e.g. Hitchcock and Simpson, 2014). This is not ~~surprising~~unexpected, as large amounts of wave drag are deposited at high altitudes, which strongly influences middle atmosphere dynamics. In the low-top models, this influence is not adequately represented. Models with more ~~levels in the vertical~~ vertical levels in the stratosphere generally perform better in our analysis. CNRM-CM6-1, which has the second-highest number of levels in the vertical, does not only have a good representation of most variables but also the best results in detecting splits and displacements. ~~The MPI-ESM1-2 ensembles on the other hand produce similar results even though they differ in~~ As stated in (Wu and Reichler, 2020), a finer vertical resolution also improves the simulation of the SSW frequency. While the models with the modest spatial (vertical and horizontal) resolution (INM-CM5-0 and MIROC6) show a good performance, especially for the aspect ratio and kurtosis at 10 hPa, the MPI-ESM1-2-LR ensemble produce better results despite its vertical and horizontal resolution. Dedicated model experiments with simulations in various horizontal and vertical resolutions are needed to systematically assess the impact of resolution on SPV representation.

An additional source of uncertainty might be the gravity wave (GW) parameterizations (e.g. Eichinger et al., 2020a; Karami et al., 2022; Eichinger et al., 2023). Events with strong gravity wave drag can affect the refractive index in the lower stratosphere (Kuchar et al., 2022). A higher refractive index results in stronger upward propagating wave activity and thus the SPV is disrupted more easily. Wu and Reichler (2020) found that the uncertainty of the refractive index in the lower stratosphere above the tropospheric jet plays an important role for the uncertainty in the simulated SSW frequency. Therefore, ~~different gravity wave parameterizations~~ these uncertainties in the models may be attributed to ~~these uncertainties~~different GW parameterizations (Sigmond and Shepherd, 2014). Recently, HÃ¡jkovÃ¡ and Å Ã¡cha (2023) showed that the SPV climatologies in CMIP6 models are largely insensitive to high latitude wave drag, but also state that SSW simulation can be sensitive to small nuances in model dynamics. Dedicated analyses are needed to fully assess the impact of various wave drag mechanisms on SPV geometry and SSWs, including consideration of the non-linear feedbacks between wave drag and mean flow. For example, Sigmond et al. (2023) have attributed the difference in simulated SSW frequency between CanESM2 (overestimation) and CanESM5 (underestimation) of the number of SSWs, especially splits as seen in Fig. 9**B**) to changes in settings of GW tuning.

Apart from model resolution and GW parameterizations and their tuning parameters, Morgenstern et al. (2022) revisited the influence of stratospheric ozone chemistry ~~influence~~ on the SPV and SSW frequency. Several additional studies demonstrated the importance of interactive ozone chemistry for representing temperature variability and extremes in the Arctic polar stratosphere (Haase and Matthes, 2019; Rieder et al., 2019; Oehrlein et al., 2020). Therefore, the way how atmospheric chemistry is treated in the model may be another factor for model skill in representing the SPV, in particular the feedback of stratospheric ozone on dynamics via radiation. CNRM-CM6-1 has a simplified but still interactive chemistry (Voldoire et al., 2019). The only analyzed model with a complete interactive chemistry is UKESM1-0-LL and overall it performs well. However, a detailed analysis of its impact on spatio-temporal SPV variability would be needed for conclusive statements.

Models that were found to simulate well the alternating easterlies and westerlies in the tropics by Richter et al. (2020) (the quasi-biennial oscillation, QBO), mostly perform better in our analysis (e.g., CNRM-CM6-1, IPSL-CM6A-LR, UKESM1-0-LL). On the other hand, models with poor QBO representation (CanESM2, CanESM5, CESM2) show a weaker performance in the rank histograms RHs and the representation of splits and displacements. The SPV is influenced via teleconnection associated with the QBO, via the so-called Holton-Tan mechanism (HTM; Holton and Tan, 1980; Baldwin et al., 2001). Rao et al. (2020) analyzed which models have a good representation of the HTM, but here we find no clear connection of a good HTM representation with a good representation of SPV variability.

While a relatively long period was regarded in the rank histogram RH analysis, it cannot be ruled out that an ensemble might show different performances during this time (Bothe et al., 2013). An option could be to analyze individual months separately, since differences in the model performance might for example occur between mid-winter, where the highest variability in wind speeds is observed, and early as well as late winter.

As has been investigated before (Hamill, 2001; Bröcker, 2008; Siegert et al., 2012), a flat rank histogram is only a necessary but not a sufficient criterion to conclude that ensemble simulations are reliable (calibrated). Due to the unknown and undersampled nature of initial-condition distributions and unavoidable simplifications and errors in the dynamical formulation, we cannot expect raw ensembles to be calibrated (**?**). This can be achieved via calibration (**??**) which can likely reduce uncertainty in climate projections (Tett et al., 2022).

Furthermore, the thresholds used for the definition of the events could be varied. Other values might lead to better resolution between steady and unsteady SPV conditions. The thresholds we used here were chosen based on the reanalysis dataset by Seviour et al. (2013) as stated in Sec. 2.5. Another important question is whether the number of ensemble members is sufficient for evaluation of the highly variable SPV representation. In particular, the INM-CM5-0, MIROC6 and MPI-ESM1-2-HR ensembles may not be large enough to fully cover effective dimension of SPV (Christiansen, 2021). This is a topic for detailed future investigation.

6 Conclusions

In this study, we assess the stratospheric polar vortex (SPV) form , stability and variability as well as the ability to distinguish different morphologies of sudden stratospheric warmings (SSWs) in large CMIP5 and CMIP6 climate model ensembles. We analyze the SPV by means of rank histograms (RHs) and the SSWs separated into splits and displacements by receiver operating characteristics curves and use ERA5 reanalysis data as reference. These analyses reveal strongly varying performances of the individual models over all SPV moment diagnostics and pressure levels. The two models that overall simulate the SPV and SSWs closest to ERA5 are CNRM-CM6-1 and UKESM1-0-LL. In contrast, results of CanESM5 and CESM2 should be handled with particular care in SPV studies, as these models did not perform well in our analysis. In general, the ensembles show a better ability in simulating displacement-type SSW events than split-type events. As SSWs represent extreme events, this model skill, however, is not always connected with representing well the geometry-based SPV diagnostics centroid latitude and aspect ratio, which diagnose SPV climatologies.

For the SPV centroid latitude and aspect ratio most ensembles are biased to some extent, but with no consistent direction among the ensembles. ~~Our~~ While regression of these ~~biases did not indicate robust~~ geometric SPV biases indicates also biases in split and displacement frequency as in Seviour et al. (2016) and Hall et al. (2021)~~. Of~~, this does not necessarily imply that bias-corrected models simulate split and displacement frequencies according to the reanalyses (Wu and Reichler, 2020). Out of all analyzed diagnostics, the kurtosis appears to be the hardest one to simulate correctly. Most of the ensembles underestimate the variability of the kurtosis. Strong biases and an underestimation of the ~~variablity~~ variability is found for the SPV area as well. Overall, this may be constituted by the difficulty of models to simulate the well-known non-linearity of stratospheric dynamics ~~(Matthewman and Esler, 2011; Cohen et al., 2014; **?**)~~ (Matthewman and Esler, 2011; Cohen et al., 2014; Eichinger et al., 2020b) and calls for caution when using these diagnostics as SSW proxies.

We ~~can conclude that~~ conclude that usually models with a higher lid and models with a finer vertical resolution generally simulate the SPV and SSWs better with reference to ERA5. However, many factors influence ~~the~~ SPV properties and ~~the~~ SSW frequency, such as interactive chemistry, gravity wave parameterizations and other dynamical processes that differ in the individual models. It is therefore not clearly assignable from this study which model characteristics are the decisive ones for representing well the SPV and its ~~behaviour. However, our findings set the basis for sensitivity experiments with individual models changing individual model components, such as vertical/horizontal resolution, gravity wave scheme, chemistry etc.~~ variability.

Knowledge of how well different climate models perform in simulating the SPV spatial variability and SSWs correctly is of utmost importance for ~~finding the adjustment screws~~ tuning and calibrating to improve their performance, as well as for assessing their reliability in future climate projections. The latter is particularly important with regard to polar stratospheric ozone and its evolution across the 21st century.

Code availability. The code that was used to produce all plots in this study is available via Zenodo (Kuchar and Öhlert, 2024).

Data availability. All processed data files for this study are provided via Mendeley Data (Kuchar, 2023).

Author contributions. AK designed the study. AK and MÖ analysed the data. MÖ and AK compiled the manuscript with inputs of all other authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This study has been supported by Deutsche Forschungsgemeinschaft under grant JA836/43-1 and within the Transregional Collaborative Research Centre SFB/TRR 172 (Project-ID 268020496), subproject D01. RE acknowledges support by the Czech Science Foundation (grant no. 21-03295S). Climate model output has been kindly provided by the US CLIVAR Working Group on Large Ensembles and by the Earth System Grid Federation (ESGF). 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 ESGF for archiving the data and providing access, and the multiple funding agencies who support CMIP6 and ESGF. For data processing, resources have been used at the Deutsches Klimarechenzentrum (DKRZ) under project ID bd1022.

References

Ayarzagüena, B., Polvani, L. M., Langematz, U., Akiyoshi, H., Bekki, S., Butchart, N., Dameris, M., Deushi, M., Hardiman, S. C., Jöckel, P., et al.: No robust evidence of future changes in major stratospheric sudden warmings: a multi-model assessment from CCMI, Atmospheric chemistry and physics, 18, 11 277–11 287, https://doi.org/10.5194/acp-18-11277-2018, 2018.

Ayarzagüena, B., Charlton-Perez, A. J., Butler, A. H., Hitchcock, P., Simpson, I. R., Polvani, L. M., Butchart, N., Gerber, E. P., Gray, L., Hassler, B., et al.: Uncertainty in the response of sudden stratospheric warmings and stratosphere-troposphere coupling to quadrupled CO2 concentrations in CMIP6 models, Journal of Geophysical Research: Atmospheres, 125, e2019JD032 345, https://doi.org/10.1029/2019JD032345, 2020.

Baldwin, M., Gray, L., Dunkerton, T., Hamilton, K., Haynes, P., Randel, W. J., Holton, J. R., Alexander, M., Hirota, I., Horinouchi, T., et al.: The quasi-biennial oscillation, Reviews of Geophysics, 39, 179–229, https://doi.org/10.1029/1999RG000073, 2001.

Baldwin, M. P. and Dunkerton, T. J.: Stratospheric harbingers of anomalous weather regimes, Science, 294, 581–584, 2001.

Baldwin, M. P., Ayarzagüena, B., Birner, T., Butchart, N., Butler, A. H., Charlton-Perez, A. J., Domeisen, D. I.I. V., Garfinkel, C. I., Garny, H., Gerber, E. P., et al.: Sudden stratospheric warmings, Reviews of Geophysics, Hegglin, M. I., Langematz, U., and Pedatella, N. M.: Sudden Stratospheric Warmings, Reviews of Geophysics, 59, e2020RG000 708, https://doi.org/https://doi.org/10.1029/2020RG000708, e2020RG000708 10.1029/2020RG000708, 2021.

Black, R. X.: Stratospheric forcing of surface climate in the Arctic Oscillation, Journal of Climate, 15, 268–277, , 2002.

Blanusa, M. L., López-Zurita, C. J., and Rasp, S.: Internal variability plays a dominant role in global climate projections of temperature and precipitation extremes, Climate Dynamics, pp. 1–15, https://doi.org/10.1007/s00382-023-06664-3, 2023.

Bothe, O., Jungclaus, J. H., Zanchettin, D., and Zorita, E.: Climate of the last millennium: ensemble consistency of simulations and reconstructions, Climate of the Past, 9, 1089–1110, https://doi.org/10.5194/cp-9-1089-2013, 2013.

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., Bekki, S., Bonnet, R., Bony, S., Bopp, L., et al.: Presentation and evaluation of the IPSL-CM6A-LR climate model, Journal of Advances in Modeling Earth Systems, 12, e2019MS002 010, https://doi.org/10.1029/2019MS002010, 2020.

Bröcker, J.: On reliability analysis of multi-categorical forecasts, Nonlinear Processes in Geophysics, 15, 661–673, , 2008.

Match]Butler2015 Butler, A. H., Seidel, D. J., Hardiman, S. C., Butchart, N., Birner, T., and Match, A.: Defining sudden stratospheric warmings, Bulletin of the American Meteorological Society, 96, 1913–1928, , 2015.

Butler, A. H., Sjoberg, J. P., Seidel, D. J., and Rosenlof, K. H.: A sudden stratospheric warming compendium, Earth System Science Data, 9, 63–76, 2017.

Charlton, A. J. and Polvani, L. M.: A new look at stratospheric sudden warmings. Part I: Climatology and modeling benchmarks, J. Climate, 20, 449–469, https://doi.org/10. 1175/JCLI3996.1, 2007.

Charlton-Perez, A. J., Baldwin, M. P., Birner, T., Black, R. X., Butler, A. H., Calvo, N., Davis, N. A., Gerber, E. P., Gillett, N., Hardiman, S., et al.: On the lack of stratospheric dynamical variability in low-top versions of the CMIP5 models, Journal of Geophysical Research: Atmospheres, 118, 2494–2505, https://doi.org/10.1002/jgrd.50125, 2013.

Christiansen, B.: The blessing of dimensionality for the analysis of climate data, Nonlinear Processes in Geophysics, 28, 409–422, 2021.

Cohen, N. Y., Gerber, E. P., and BühlerBÃ¼hler, O.: What Drives the Brewer–Brewerâ“Dobson Circulation?, Journal of the Atmospheric Sciences, 71, 3837 – 3855, https://doi.org/https://doi.org/10.1175/JAS-D-14-0021.1, 2014.

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D., DuVivier, A., Edwards, J., Emmons, L., Fasullo, J., Garcia, R., Gettelman, A., et al.: The community earth system model version 2 (CESM2), Journal of Advances in Modeling Earth Systems, 12, e2019MS001 916, https://doi.org/10.1029/2019MS001916, 2020.

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., Fiore, A., Frankignoul, C., Fyfe, J. C., Horton, D. E., et al.: Insights from Earth system model initial-condition large ensembles and future prospects, Nature Climate Change, 10, 277–286, https://doi.org/10.1038/s41558-020-0731-2, 2020.

Donner, L. J., Wyman, B. L., Hemler, R. S., Horowitz, L. W., Ming, Y., Zhao, M., Golaz, J.-C., Ginoux, P., Lin, S.-J., Schwarzkopf, M. D., et al.: The dynamical core, physical parameterizations, and basic simulation characteristics of the atmospheric component AM3 of the GFDL global coupled model CM3, Journal of Climate, 24, 3484–3519, https://doi.org/10.1175/2011JCLI3964.1, 2011.

Eichinger, R., Garny, H., Šácha, P., Danker, J., Dietmüller, S., and Oberländer-Hayn, S.: Effects of missing gravity waves on stratospheric dynamics; part 1: climatology, Climate Dynamics, 54, 3165–3183, 2020. 2020a.

Eichinger, R., Garny, H., Šácha, P., Danker, J., Dietmüller, S., and Oberländer-Hayn, S.: Effects of missing gravity waves on stratospheric dynamics; part 1: climatology, Climate Dynamics, 54, 3165–3183, 2020b.

Eichinger, R., Rhode, S., Garny, H., Preusse, P., Pisoft, P., Kuchar, A., Jöckel, P., Kerkweg, A., and Kern, B.: Emulating lateral gravity wave propagation in a global chemistry-climate model (EMAC v2.55.2) through horizontal flux redistribution, EGUsphere, 2023, 1–34, https://doi.org/10.5194/egusphere-2023-270, 2023.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geoscientific Model Development, 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

Gerber, E., Martineau, P., Ayarzaguena, B., Barriopedro, D., Bracegirdle, T., Butler, A., Calvo, N., Hardiman, S., Hitchcock, P., Iza, M., et al.: Extratropical stratosphere–troposphere coupling, SPARC Reanalysis Intercomparison Project (S-RIP) Final Report, 2022.

Haase, S. and Matthes, K.: The importance of interactive chemistry for stratosphere–troposphere coupling, Atmospheric Chemistry and Physics, 19, 3417–3432, https://doi.org/10.5194/acp-19-3417-2019, 2019.

Hall, R. J., Mitchell, D. M., Seviour, W. J., and Wright, C. J.: Persistent model biases in the CMIP6 representation of stratospheric polar vortex variability, Journal of Geophysical Research: Atmospheres, 126, e2021JD034 759, https://doi.org/10.1029/2021JD034759, 2021.

Hamill, T. M.: Interpretation of rank histograms for verifying ensemble forecasts, Monthly Weather Review, 129, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2, 2001.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hitchcock, P. and Simpson, I. R.: The Downward Influence of Stratospheric Sudden Warmings, Journal of the Atmospheric Sciences, 71, 3856 – 3876, https://doi.org/https://doi.org/10.1175/JAS-D-14-0012.1, 2014.

Hoffmann, L. and Spang, R.: An assessment of tropopause characteristics of the ERA5 and ERA-Interim meteorological reanalyses, Atmospheric Chemistry and Physics, 22, 4019–4046, https://doi.org/10.5194/acp-22-4019-2022, 2022.

Holton, J. R.: The Dynamics of Sudden Stratospheric Warmings, Annual Review of Earth and Planetary Sciences, 8, 169–190, https://doi.org/10.1146/annurev.ea.08.050180.001125, 1980.

Holton, J. R. and Tan, H.-C.: The influence of the equatorial quasi-biennial oscillation on the global circulation at 50 mb, Journal of Atmospheric Sciences, 37, 2200–2208, https://doi.org/10.1175/1520-0469(1980)037<2200:TIOTEQ>2.0.CO;2, 1980.

HÃ¡jkovÃ¡, D. and Å Ã¡cha, P.: Parameterized orographic gravity wave drag and dynamical effects in CMIP6 models, Climate Dynamics, https://doi.org/10.1007/s00382-023-07021-0, 2023.

Jolliffe, I. T. and Primo, C.: Evaluating rank histograms using decompositions of the chi-square test statistic, Monthly Weather Review, 136, 2133–2139, https://doi.org/10.1175/2007MWR2219.1, 2008.

Jucker, M., Reichler, T., and Waugh, D. W.: How frequent are Antarctic sudden stratospheric warmings in present and future climate?, Geophysical Research Letters, 48, e2021GL093 215, https://doi.org/10.1029/2021GL093215, 2021.

Karami, K., Mehrdad, S., and Jacobi, C.: Response of the resolved planetary wave activity and amplitude to turned off gravity waves in the UA-ICON general circulation model, Journal of Atmospheric and Solar-Terrestrial Physics, 241, 105 967, https://doi.org/https://doi.org/10.1016/j.jastp.2022.105967, 2022.

Kharin, V. V. and Zwiers, F. W.: On the ROC Score of Probability Forecasts, Journal of Climate, 16, 4145 – 4150, , 2003.

Baldwin, and Gray]Kidston2015 Kidston, J., Scaife, A. A., Hardiman, S. C., Mitchell, D. M., Butchart, N., Baldwin, M. P., and Gray, L. J.: Stratospheric influence on tropospheric jet streams, storm tracks and surface weather, Nature Geoscience, 8, 433–440, , 2015.

King, A. D., Butler, A. H., Jucker, M., Earl, N. O., and Rudeva, I.: Observed relationships between sudden stratospheric warmings and European climate extremes, Journal of Geophysical Research: Atmospheres, 124, 13 943–13 961, https://doi.org/10.1029/2019JD030480, 2019.

Kirchmeier-Young, M. C., Zwiers, F. W., and Gillett, N. P.: Attribution of extreme events in Arctic sea ice extent, Journal of Climate, 30, 553–571, https://doi.org/10.1175/JCLI-D-16-0412.1, 2017.

Kuchar, A.: Accompanying data to "On the reliability of large ensembles simulating the Northern Hemispheric winter stratospheric polar vortex", Mendeley Data, https://doi.org/10.17632/d6yg8ncppg.1, 2023.

Kuchar, A. and Öhlert, M.: VACILT/reliability_LE: Third Fourth release of our code repository related to reliability of large ensembles, , 2023. https://doi.org/10.5281/zenodo.10620326, 2024.

Kuchar, A., Sacha, P., Eichinger, R., Jacobi, C., Pisoft, P., and Rieder, H.: On the impact of Himalaya-induced gravity waves on the polar vortex, Rossby wave activity and ozone, EGUsphere, pp. 1–22, https://doi.org/https://doi.org/10.5194/egusphere-2022-474, 2022.

Langematz, U., Meul, S., Grunow, K., Romanowsky, E., Oberländer, S., Abalichin, J., and Kubin, A.: Future Arctic temperature and ozone: The role of stratospheric composition changes, Journal of Geophysical Research: Atmospheres, 119, 2092–2112, https://doi.org/10.1002/2013JD021100, 2014.

Lawrence, Z. D., Perlwitz, J., Butler, A. H., Manney, G. L., Newman, P. A., Lee, S. H., and Nash, E. R.: The remarkably strong Arctic stratospheric polar vortex of winter 2020: Links to record-breaking Arctic oscillation and ozone loss, Journal of Geophysical Research: Atmospheres, 125, e2020JD033 271, https://doi.org/10.1029/2020JD033271, 2020.

Maher, N., Milinski, S., Suarez-Gutierrez, L., Botzet, M., Dobrynin, M., Kornblueh, L., Kröger, J., Takano, Y., Ghosh, R., Hedemann, C., et al.: The Max Planck Institute Grand Ensemble: enabling the exploration of climate system variability, Journal of Advances in Modeling Earth Systems, 11, 2050–2069, , 2019.

Manzini, E., Karpechko, A. Y., Anstey, J., Baldwin, M., Black, R., Cagnazzo, C., Calvo, N., Charlton-Perez, A., Christiansen, B., Davini, P., et al.: Northern winter climate change: Assessment of uncertainty in CMIP5 projections related to stratosphere-troposphere coupling, Journal of Geophysical Research: Atmospheres, 119, 7979–7998, https://doi.org/10.1002/2013JD021403, 2014.

Matthewman, N. J. and Esler, J. G.: Stratospheric Sudden Warmings as Self-Tuning Resonances. Part I: Vortex Splitting Events, Journal of the Atmospheric Sciences, 68, 2481 – 2504, https://doi.org/https://doi.org/10.1175/JAS-D-11-07.1, 2011.

Matthewman, N. J., Esler, J. G., Charlton-Perez, A. J., and Polvani, L. M.: A new look at stratospheric sudden warmings. Part III: Polar vortex evolution and vertical structure, Journal of Climate, 22, 1566–1585, https://doi.org/10.1175/2008JCLI2365.1, 2009.

Maycock, A. C. and Hitchcock, P.: Do split and displacement sudden stratospheric warmings have different annular mode signatures?, Geophysical Research Letters, 42, 10,943–10,951, https://doi.org/https://doi.org/10.1002/2015GL066754, 2015.

Mitchell, D. M., Charlton-Perez, A. J., and Gray, L. J.: Characterizing the Variability and Extremes of the Stratospheric Polar Vortices Using 2D Moment Analysis, Journal of the Atmospheric Sciences, 68, 1194 – 1213, https://doi.org/10.1175/2010JAS3555.1, 2011.

Mitchell, D. M., Gray, L. J., Anstey, J., Baldwin, M. P., and Charlton-Perez, A. J.: The influence of stratospheric vortex displacements and splits on surface climate, Journal of climate, 26, 2668–2682, 2013.

Morgenstern, O., Kinnison, D. E., Mills, M., Michou, M., Horowitz, L. W., Lin, P., Deushi, M., Yoshida, K., O’Oâ™Connor, F. M., Tang, Y., et al.: Comparison of Arctic and Antarctic Stratospheric Climates in Chemistry Versus No-Chemistry Climate Models, Journal of Geophysical Research: Atmospheres, 127, e2022JD037 123, https://doi.org/10.1029/2022JD037123, 2022.

Müller, W. A., Jungclaus, J. H., Mauritsen, T., Baehr, J., Bittner, M., Budich, R., Bunzel, F., Esch, M., Ghosh, R., Haak, H., et al.: A higher-resolution version of the max planck institute earth system model (MPI-ESM1. 2-HR), Journal of Advances in Modeling Earth Systems, 10, 1383–1413, 2018.

Oehrlein, J., Chiodo, G., and Polvani, L. M.: The effect of interactive ozone chemistry on weak and strong stratospheric polar vortex events, Atmospheric Chemistry and Physics, 20, 10 531–10 544, https://doi.org/10.5194/acp-20-10531-2020, 2020.

Olonscheck, D., Suarez-Gutierrez, L., Milinski, S., Beobide-Arsuaga, G., Baehr, J., FrÃ¶b, F., Ilyina, T., Kadow, C., Krieger, D., Li, H., Marotzke, J., PlÃ©siat, Ã., Schupfner, M., Wachsmann, F., Wallberg, L., Wieners, K.-H., and Brune, S.: The New Max Planck Institute Grand Ensemble With CMIP6 Forcing and High-Frequency Model Output, Journal of Advances in Modeling Earth Systems, 15, e2023MS003 790, https://doi.org/https://doi.org/10.1029/2023MS003790, e2023MS003790 2023MS003790, 2023.

Rao, J. and Garfinkel, C. I.: CMIP5/6 models project little change in the statistical characteristics of sudden stratospheric warmings in the 21st century, Environmental Research Letters, 16, 034 024, https://doi.org/10.1088/1748-9326/abd4fe, 2021.

Rao, J., Garfinkel, C. I., and White, I. P.: How does the quasi-biennial oscillation affect the boreal winter tropospheric circulation in CMIP5/6 models?, Journal of Climate, 33, 8975–8996, https://doi.org/10.1175/JCLI-D-20-0024.1, 2020.

Richter, J. H., Anstey, J. A., Butchart, N., Kawatani, Y., Meehl, G. A., Osprey, S., and Simpson, I. R.: Progress in simulating the quasi-biennial oscillation in CMIP models, Journal of Geophysical Research: Atmospheres, 125, e2019JD032 362, https://doi.org/10.1029/2019JD032362, 2020.

Rieder, H. E., Chiodo, G., Fritzer, J., Wienerroither, C., and Polvani, L. M.: Is interactive ozone chemistry important to represent polar cap stratospheric temperature variability in Earth-System Models?, Environmental Research Letters, 14, 044 026, https://doi.org/10.1088/1748-9326/ab07ff, 2019.

Scaife, A. A., Knight, J. R., Vallis, G. K., and Folland, C. K.: A stratospheric influence on the winter NAO and North Atlantic surface climate, Geophysical Research Letters, 32, , 2005.

Scaife, A. A., Spangehl, T., Fereday, D. R., Cubasch, U., Langematz, U., Akiyoshi, H., Bekki, S., Braesicke, P., Butchart, N., Chipperfield, M. P., et al.: Climate change projections and stratosphere–troposphere interaction, Climate Dynamics, 38, 2089–2097, https://doi.org/10.1007/s00382-011-1080-7, 2012.

Sellar, A. A., Jones, C. G., Mulcahy, J. P., Tang, Y., Yool, A., Wiltshire, A., O'connor, F. M., Stringer, M., Hill, R., Palmieri, J., et al.: UKESM1: Description and evaluation of the UK Earth System Model, Journal of Advances in Modeling Earth Systems, 11, 4513–4558, https://doi.org/10.1029/2019MS001739, 2019.

Seviour, W. J., Mitchell, D. M., and Gray, L. J.: A practical method to identify displaced and split stratospheric polar vortex events, Geophysical Research Letters, 40, 5268–5273, 2013.

Seviour, W. J., Gray, L. J., and Mitchell, D. M.: Stratospheric polar vortex splits and displacements in the high-top CMIP5 climate models, Journal of Geophysical Research: Atmospheres, 121, 1400–1413, 2016.

Siegert, S., Bröcker, J., and Kantz, H.: Rank Histograms of Stratified Monte Carlo Ensembles, Monthly Weather Review, 140, 1558 – 1571, , 2012.

Shiogama, H., Tatebe, H., Hayashi, M., Abe, M., Arai, M., Koyama, H., Imada, Y., Kosaka, Y., Ogura, T., and Watanabe, M.: MIROC6 Large Ensemble (MIROC6-LE): experimental design and initial analyses, Earth System Dynamics, 14, 1107–1124, https://doi.org/10.5194/esd-14-1107-2023, 2023.

Sigmond, M. and Shepherd, T. G.: Compensation between Resolved Wave Driving and Parameterized Orographic Gravity Wave Driving of the Brewer–Brewerâ"Dobson Circulation and Its Response to Climate Change, Journal of Climate, 27, 5601–5610, https://doi.org/10.1175/JCLI-D-13-00644.1, 2014.

Sigmond, M., ScinoccaAnstey, J., Arora, V., Digby, R., Gillett, N., Kharin, V., Merryfield, W., Reader, C., Scinocca, J., Swart, N., Virgin, J., Abraham, C., Cole, J., Lambert, N., Lee, W.-S., Liang, Y., Malinina, E., Rieger, L., von Salzen, K., Seiler, C., Seinen, C., Shao, A., Sospedra-Alfonso, R., Wang, L., and Yang, D.: Improvements in the Canadian Earth System Model (CanESM) through systematic model analysis: CanESM5.0 and Shepherd, T.: Enhanced seasonal forecast skill following stratospheric sudden warmings, Nature Geoscience, 6, 98–102, , 2013. CanESM5.1, Geoscientific Model Development, 16, 6553–6591, https://doi.org/10.5194/gmd-16-6553-2023, 2023.

Simpson, I. R., Hitchcock, P., Seager, R., Wu, Y., and Callaghan, P.: The downward influence of uncertainty in the Northern Hemisphere stratospheric polar vortex response to climate change, Journal of Climate, 31, 6371–6391, https://doi.org/10.1175/JCLI-D-18-0041.1, 2018.

Suarez-Gutierrez, L., Milinski, S., and Maher, N.: Exploiting large ensembles for a better yet simpler climate model evaluation, Climate Dynamics, 57, 2557–2580, https://doi.org/10.1007/s00382-021-05821-w, 2021.

Swart, N. C., Cole, J. N., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., Anstey, J., Arora, V., Christian, J. R., Hanna, S., et al.: The Canadian earth system model version 5 (CanESM5. 0.3), Geoscientific Model Development, 12, 4823–4873, https://doi.org/10.5194/gmd-12-4823-2019, 2019.

Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogochi, K., Takemura, T., Sudo, K., Sekiguchi, M., Abe, M., Saito, F., et al.: Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6, Geoscientific Model Development, 12, 2727–2765, https://doi.org/10.5194/gmd-12-2727-2019, 2019.

Tett, S. F. B., Gregory, J. M., Freychet, N., Cartis, C., Mineter, M. J., and Roberts, L.: Does Model Calibration Reduce Uncertainty in Climate Projections?, Journal of Climate, 35, 2585 – 2602, , 2022.

Thompson, D. W., Baldwin, M. P., and Wallace, J. M.: Stratospheric connection to Northern Hemisphere wintertime weather: Implications for prediction, Journal of Climate, 15, 1421–1428, 2002.

Tripathi, O. P., Baldwin, M., Charlton-Perez, A., Charron, M., Eckermann, S. D., Gerber, E., Harrison, R. G., Jackson, D. R., Kim, B.-M., Kuroda, Y., et al.: The predictability of the extratropical stratosphere on monthly time-scales and its impact on the skill of tropospheric forecasts, Quarterly Journal of the Royal Meteorological Society, 141, 987–1003, , 2015.

Messner]vannitsem2018statistical Vannitsem, S., Wilks, D. S., and Messner, J.: Statistical postprocessing of ensemble forecasts, Elsevier, 2018.

Vokhmyanin, M., Asikainen, T., Salminen, A., and Mursula, K.: Long-Term Prediction of Sudden Stratospheric Warmings With Geomagnetic and Solar Activity, Journal of Geophysical Research: Atmospheres, 128, e2022JD037 337, https://doi.org/https://doi.org/10.1029/2022JD037337, e2022JD037337 2022JD037337, 2023.

Voldoire, A., Saint-Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M., Colin, J., Guérémy, J.-F., Michou, M., Moine, M.-P., et al.: Evaluation of CMIP6 deck experiments with CNRM-CM6-1, Journal of Advances in Modeling Earth Systems, 11, 2177–2213, https://doi.org/10.1029/2019MS001683, 2019.

Volodin, E. and Gritsun, A.: Simulation of observed climate changes in 1850–2014 with climate model INM-CM5, Earth System Dynamics, 9, 1235–1242, https://doi.org/10.5194/esd-9-1235-2018, 2018.

Wilks, D. S.: Statistical methods in the atmospheric sciences, vol. 100, Academic press, 2011.

Wilks, D. S.: Enforcing calibration in ensemble postprocessing, Quarterly Journal of the Royal Meteorological Society, 144, 76–84, , 2018.

Wilks, D. S.: Indices of Rank Histogram Flatness and Their Sampling Properties, Monthly Weather Review, 147, 763 – 769, , 2019.

Wu, Z. and Reichler, T.: Variations in the frequency of stratospheric sudden warmings in CMIP5 and CMIP6 and possible causes, Journal of Climate, 33, 10 305–10 320, https://doi.org/10.1175/JCLI-D-20-0104.1, 2020.

Zappa, G. and Shepherd, T. G.: Storylines of atmospheric circulation change for European regional climate impact assessment, Journal of Climate, 30, 6561–6577, https://doi.org/10.1175/JCLI-D-16-0807.1, 2017.