ABSTRACT: The climatological frequency of stratospheric sudden warming events (SSWs) is an important dynamical characteristic of the extratropical stratosphere. However, modern climate models have difficulties in simulating this frequency, with many models either considerably under- or overestimating the observational estimates. Past research has found that models with a higher upper lid tend to simulate a higher and more realistic number of SSWs. The present study revisits this issue and investigates causes for biases in the simulated SSW frequency from the CMIP5 and CMIP6 models. It is found that variations in the frequency are closely related to 1) the strength of the polar vortex and 2) the upward-propagating wave activity in the stratosphere. While it is difficult to explain the variations in the polar vortex strength from the available model output, the stratospheric wave activity is influenced by different aspects of the climatological mean state of the atmosphere in the lower stratosphere. We further find that models with a finer vertical resolution in the stratosphere are overall more realistic: vertical resolution is associated with a smaller cold bias above the extratropical tropopause, more upward-propagating wave activity in the lower stratosphere, and a higher frequency of SSWs. We conclude that not only a high model lid but also a fine vertical resolution in the stratosphere is important for simulating the dynamical variability of the stratosphere.

KEYWORDS: Atmospheric circulation; Stratosphere-troposphere coupling; Stratosphere; Climate models; Model evaluation/performance; Model output statistics

1. Introduction

Stratospheric sudden warming events (SSWs) represent extreme perturbations of the wintertime Arctic stratosphere and contribute significantly to the predictability of tropospheric weather (Sigmond et al. 2013; Tripathi et al. 2015; Karpechko et al. 2017). Observations show that SSWs occur on average in 6 out of 10 years, a frequency that is an essential characteristic of the coupled stratosphere–troposphere system (Butler et al. 2019). Given this fundamental nature, it is surprising that even the most sophisticated models used for climate assessments have major difficulties in reproducing this frequency. This is demonstrated in Fig. 1, comparing the SSW frequencies of models from phases 5 and 6 of the Coupled Model Intercomparison Projects (CMIP5 and CMIP6). The simulated frequencies vary widely, with some models producing no or almost no SSWs, and others simulating more than one SSW per year. Similar variations have been reported from other model intercomparison studies (Charlton et al. 2007; SPARC 2010; Butchart et al. 2011; Charlton-Perez et al. 2013; Ayarzagüena et al. 2018, 2020). With this in mind, one may ask why models have such difficulties in reproducing the observed number of SSWs, and which aspect of a model needs to be improved to correct the situation. Answering this question and reducing the SSW biases of models may lead to better climate predictions, as there is currently neither robust evidence nor a clear consensus on how the number of SSWs will respond to climate change (Ayarzagüena et al. 2018).

One difficulty in reproducing the observed SSW frequency is the relatively rare nature of the events, creating large inter-annual variability and necessitating relatively long time series to achieve statistically representative results. However, there is much more to simulating a realistic SSW frequency. For example, SSWs depend on the upward propagation of planetary waves into the polar vortex region, and the interaction of the waves with the mean vortex winds. The wave propagation is sensitive to subtle variations in the stratospheric base state, as expressed by the index of refraction for planetary waves (Matsuno 1970). The index contains higher derivatives of winds and temperatures, which are difficult to simulate correctly. The model generated spectrum of planetary waves, model resolution, and uncertainties about the parameterization of gravity wave drag only add to the list of complicating factors.

Charlton-Perez et al. (2013) and other studies have shown that models with a low upper lid tend to underestimate the SSW frequency. Lee and Black (2015) indicated that a low model top underestimates the variability of the stratospheric planetary wave activity and of the polar vortex strength. Shaw and Perlwitz (2010) demonstrated that a low-top model creates excessive wave damping near the upper boundary, reduced extreme heat flux events, and too few SSWs. However, a high model top alone does not guarantee a realistic SSW frequency (e.g., the two GFDL models in Fig. 1), and some models with a low lid simulate reasonable frequencies (e.g., CNRM-CM5 and BCC-ESM1 in Fig. 1).
It has also been suggested that the simulated SSW frequency is linked to climatological characteristics of a model. Examples for such characteristics include the meridional heat flux in the lower stratosphere (Charlton et al. 2007) and the strength of the polar vortex (Charlton-Perez et al. 2013). Horan and Reichler (2017) demonstrated that the seasonal variations in the SSW frequency can be accurately explained from the seasonal cycle statistics of the polar vortex strength (i.e., its mean, standard deviation, and skewness). Similarly, Taguchi (2017) related the intermodel spread of the SSW frequency in CMIP5 to the polar vortex strength and its variability.

The region just above the extratropical tropopause, the so-called upper troposphere–lower stratosphere (UTLS) region, where the vertical profiles of temperature, wind, and chemical constituents change dramatically (Gerber and Manzini 2016), is also relevant for SSWs. For example, Chen and Robinson (1992) highlighted the important role of the tropopause in regulating the upward propagation of planetary waves into the stratosphere. Various other studies came to similar conclusions (Shaw et al. 2014; Jucker 2016; de la Cámara et al. 2017; Martineau et al. 2018). Sjoberg and Birner (2014) suggested that the tropopause inversion layer could potentially be a source of wave activity in the lower stratosphere, and Birner and Albers (2017) found that the tropopause layer is crucial for a nonlinear wave–mean flow feedback that can cause SSWs. There is also consensus that sufficient vertical resolution is one of the essential criteria to correctly locate the tropopause and represent the dynamical processes in its vicinity (Birner 2006; Hegglin et al. 2010; Gettelman et al. 2011; Gerber and Manzini 2016). In a recent study, Wang et al. (2019) showed how increased vertical resolution improves the representation of the tropical tropopause in the WACCM model.

The goal of the present study is to revisit the SSW frequency problem and identify some of the processes that are involved in the simulation of the correct frequency. This is accomplished by examining the CMIP5 and CMIP6 ensembles and relating their intermodel spread in SSW frequency to variations in other basic quantities. As we will show, the strength of the polar vortex is the most important indicator for the number of SSWs, but the simulation of the climatological mean state in
the UTLS region is also important. We further find that a finer vertical model resolution in the stratosphere tends to improve the simulation of the climatological mean state and the upward wave activity flux. Last, this study also sheds some new light on the performance of the two latest generations of climate models in the stratosphere.

The paper is organized as follows. Section 2 describes our data and diagnostic methods. The main results are described in section 3, identifying how the model simulated SSW frequency is related to other climatological characteristics, and explaining the underlying dynamical mechanisms. The paper ends with section 4, providing a summary, a discussion, and conclusions.

2. Models and analysis

a. Models

The data for this study consist of daily and monthly variables from the historical experiments of the CMIP5 (1950–2005) and CMIP6 (1950–2014) projects. As shown in Fig. 1, we investigate 23 CMIP5 models and 20 CMIP6 models, based on the availability of daily model output. We use only one ensemble member for each model, which is in most cases r1i1p1f1 (CMIP5) or r1i1p1f1 (CMIP6). Our analysis is carried out for November–March (NDJFM) since this is the time when SSWs are defined. The model climatologies are validated against ERA-40 (Uppala et al. 2005), which covers a similar time period (1958–2001) as the model simulations. We also use the more modern reanalysis ERA-Interim (1979–2016; Dee et al. 2011) to repeat some of our analysis, leading to very similar results; we therefore present our results only for ERA-40. To define SSWs and calculate the Eliassen–Palm (EP) fluxes (Eliassen and Palm 1961), we employ daily output of threedimensional (longitude, latitude, level) zonal wind, meridional wind, and temperature. The daily output from the CMIP models, however, is only archived at 8 vertical pressure levels (1000, 850, 700, 500, 250, 100, 50, 10 hPa). This somewhat limits our ability to calculate reliable EP fluxes in the stratosphere, because doing so involves taking vertical derivatives. The monthly model output, however, is available for at least 17 pressure levels, with 9 levels (250, 200, 150, 100, 70, 50, 30, 20, 10 hPa) in the stratosphere. We therefore calculate quantities that do not require daily data from monthly data.

b. Analysis

Our SSW definition follows the criterion of Charlton and Polvani (2007), based on the reversal of the daily zonal-mean zonal wind at 10 hPa and 60°N (U1000), and a return to westerlies afterward for at least 10 consecutive days. Two wind reversals in the same season are treated as distinct SSWs if they are separated by at least 20 days. The SSW frequency (fSSW) is the number of SSWs per 100 years. Based on ERA-40 over the period 1958–2001, the observation-based climatological SSW frequency is 61% (Butler et al. 2017), or in other words, the probability of an SSW in any given year is ρ = 0.61. In calculating a confidence interval for ρ, we apply a nonparametric bootstrapping method similar to Ayarzagüena et al. (2020). We perform the bootstrapping on individual years with replacement, using the same number of years as the actual number of years. We repeat 50,000 times, compute each time fSSW, and derive the 2.5th–97.5th-percentile range of these samples. The 95% confidence interval of fSSW in ERA-40 is 43%–80%. The intervals for the models are calculated in the same way.

SSWs are defined using the zero-wind threshold, and one reason for biases in crossing this threshold is the simulated climatological strength of the vortex. To estimate the contribution from this, we follow Scaife et al. (2010) and correct a model’s climatological vortex strength by the reanalysis climatological U1000 but retain the variability of the vortex. We then use the corrected U1000 time series to detect SSWs and derive a new SSW frequency.

We employ EP flux diagnostics to analyze the wave–mean flow interaction. More specifically, we calculate the horizontal and vertical components of the quasigeostrophic version of the EP flux vector F (Edmon et al. 1981) in pressure coordinates using daily data. A complicating issue arises when calculating the EP fluxes from vertically coarsely resolved model output. To assess the impact of coarse vertical resolution, we use daily ERA-40 data to calculate the EP fluxes from using the same pressure levels that are available for the daily model data. The result is then compared to the EP fluxes from using the original data with all pressure levels. We find that the vertical component of the EP flux (FZ) from the vertically reduced data consistently underestimates the fluxes compared to using data from all levels. To further examine the influence of coarse vertical resolution on the model-to-model variation in CMIP5, we calculate the EP fluxes from monthly data twice, first, by using all levels, and second, by using only a subset of levels. The results suggest that the variations of FZ across the 23 models calculated from using 1) sublevels and 2) all levels are very similar. To calculate the upward wave activity, we could have also replaced the vertical EP fluxes with the poleward eddy heat fluxes, which do not necessitate calculating vertical derivatives. We repeated our analysis using heat fluxes and found very similar results. However, Taguchi (2017) suggested that FZ is more appropriate than heat fluxes in a multimodel comparison because variations in static stability are also important for the propagation of the waves. We therefore use FZ to diagnose the waves and their forcing.

We use the 40°–70°N latitude-weighted climatological FZ at 100 hPa (FZ100) for the strength of the stratospheric wave driving, and the zonal-mean zonal wind at 10 hPa and 60°N (U1000) for the strength of the stratospheric polar vortex. The propagation of planetary waves depends on the atmospheric refractive properties, expressed in terms of an unitless quasigeostrophic index of refraction squared (R^2) (Matsumo 1970). Assuming the planetary waves are stationary, R^2 is given by

\[ R^2 = \frac{U^2}{g_\nu} - \frac{k^2}{\cos^2 \varphi} - \frac{a f}{2 \mu_N} \left( \frac{\alpha f}{2 \mu_N} \right)^2, \]  

(1)

where

\[ \frac{1}{\mu_N} = \frac{1}{\mu_N} \frac{\partial \Phi}{\partial \varphi} = \beta - \frac{1}{\mu_N} \frac{\partial}{\partial \varphi} \left[ \frac{1}{\cos \varphi} \frac{\partial U \cos \varphi}{\partial \phi} \right] \]

\[ + \frac{f^2}{H \mu_N} \frac{\partial U}{\partial z} - \frac{f^2}{N^2} \frac{\partial U}{\partial z} - \frac{f^2}{N^2} \frac{\partial U}{\partial z} \frac{1}{N^2} \]  

(2)
is the meridional gradient of potential vorticity, $a$ is the radius of Earth, $k = 1$ is the zonal wavenumber of the wave, $\phi$ is latitude, $f$ is the Coriolis parameter, $H$ is the scale height, and $N$ is the buoyancy frequency. The thermally defined tropopause pressure is calculated from three-dimensional monthly temperature fields, following the algorithm of Reichler et al. (2003).

We follow Charlton-Perez et al. (2013) and use the pressure of the highest model level to classify the models into high-top (above 1 hPa) and low-top (below 1 hPa) models. An asterisk after the model name in Fig. 1 indicates models with a high top. With this definition, 10 out of 23 CMIP5 models and 13 out of 20 CMIP6 models are high-top models. The mean vertical resolution in the stratosphere ($r_{vertical}$) of each model (Table 1) is derived by dividing the vertical distance between 100 and 1 hPa (or the model top, whichever is lower) by the number of native model levels in this layer. We also calculate the vertical resolution only in the 300–100-hPa layer (not shown) and find that in both ensembles it is strongly correlated with $r_{vertical}$ ($r = 0.7$ in CMIP5 and 0.8 in CMIP6).

Throughout our study, we use linear regression to explain variations in $f_{SSW}$ from the intermodel spread in some other climatological quantities. To be more specific, we first use a one-predictor (unary) linear regression to link intermodel variations in $f_{SSW}$ to one quantity. Then, we use bilinear regression to estimate $f_{SSW}$ from the original quantity and one additional quantity. From this, we calculate the increase in explained variance. We use bootstrapping to determine the statistical significance of the explained variance and its increase by choosing 50 000 random model combinations with replacement, with each combination containing the same number of models as the original sample. The regressions reported in this study are all statistically significant at the 95% limit, unless otherwise specified.

3. Results

Figure 1 shows the monthly and seasonal distribution of the SSW frequency of the CMIP5 and CMIP6 models, their multimodel means (MMM), ERA-40, and ERA-Interim. Overall, there is a large spread in $f_{SSW}$ among both groups of models, ranging from 0% to more than 100%. The mean $f_{SSW}$ is 37% in CMIP5 and 54% in CMIP6, with a standard deviation of 29% and 25%, respectively. This can be compared to $f_{SSW} = 61\% \pm 18\%$ in ERA-40 (1958–2001) and $f_{SSW} = 63\% \pm 21\%$ in ERA-Interim (1979–2016). The models’ root-mean-square (rms) error with respect to the ERA-40 is 37% (CMIP5) and 26% (CMIP6). Just considering the high-top models from each ensemble reduces the rms error to 24% in CMIP5 and 17% in CMIP6. If consistency with ERA-40 is measured by the overlap of a model’s confidence interval with that of ERA-40 (gray bar), then 48% of the CMIP5 and 75% of the CMIP6 models fall into this category. Therefore, CMIP6 is closer to ERA-40, with fewer extreme outliers compared to CMIP5. In particular, more than two-thirds of the CMIP5 models have a lower frequency than ERA-40. Models with a high overall $f_{SSW}$ tend to produce more early events, in November (purple) and December (blue). In ERA-40, most SSWs occur in January (orange), while the models produce most events later in winter, consistent with Horan and Reichler (2017). For example, in the MMM, March events (green) comprise 33% of all SSWs in CMIP5 and 29% in CMIP6, whereas in ERA-40 March events amount only to 19%.

<table>
<thead>
<tr>
<th>CMIP5</th>
<th>$r_{vertical}$ (m)</th>
<th>CMIP6</th>
<th>$r_{vertical}$ (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMCC-CMS</td>
<td>733</td>
<td>MPI-ESM1–2-HR</td>
<td>768</td>
</tr>
<tr>
<td>MIROC-ESM</td>
<td>750</td>
<td>MIROC6</td>
<td>871</td>
</tr>
<tr>
<td>HadGEM2-CC</td>
<td>1900</td>
<td>MRI-ESM2–0</td>
<td>1075</td>
</tr>
<tr>
<td>MPI-ESM-LR</td>
<td>2015</td>
<td>EC-Earth3</td>
<td>1112</td>
</tr>
<tr>
<td>CMCC-CESM</td>
<td>2015</td>
<td>EC-Earth3-Veg</td>
<td>1112</td>
</tr>
<tr>
<td>MIROC5</td>
<td>2065</td>
<td>CNRM-CM6–1</td>
<td>1112</td>
</tr>
<tr>
<td>GFDL CM3</td>
<td>2149</td>
<td>CNRM-ESM2–1</td>
<td>1112</td>
</tr>
<tr>
<td>ACCESS1.0</td>
<td>2238</td>
<td>UKESM1–0-LL</td>
<td>1151</td>
</tr>
<tr>
<td>ACCESS1.3</td>
<td>2238</td>
<td>HadGEM3–GC31-LL</td>
<td>1151</td>
</tr>
<tr>
<td>IPSL-CM5A-LR</td>
<td>2303</td>
<td>IPSL-CM6A-LR</td>
<td>1273</td>
</tr>
<tr>
<td>IPSL-CM5A-MR</td>
<td>2303</td>
<td>BCC-CSM2-MR</td>
<td>1345</td>
</tr>
<tr>
<td>MRI-CGCM3</td>
<td>2303</td>
<td>CESM2-WACCM</td>
<td>1600</td>
</tr>
<tr>
<td>INM-CM4</td>
<td>2686</td>
<td>CanESM5</td>
<td>1639</td>
</tr>
<tr>
<td>BCC-CSM1.1-m</td>
<td>3093</td>
<td>NorESM2-LM</td>
<td>2455</td>
</tr>
<tr>
<td>BNU-ESM</td>
<td>3093</td>
<td>GISS-E2–1-G</td>
<td>2931</td>
</tr>
<tr>
<td>CMCC-CM</td>
<td>3224</td>
<td>FGOALS-f3-L</td>
<td>2983</td>
</tr>
<tr>
<td>CNRM-CM5</td>
<td>3224</td>
<td>SAM0-UNICON</td>
<td>3000</td>
</tr>
<tr>
<td>CCSR4</td>
<td>3342</td>
<td>BCC-ESM1</td>
<td>3093</td>
</tr>
<tr>
<td>NorESM1-M</td>
<td>3342</td>
<td>CESM2</td>
<td>3300</td>
</tr>
<tr>
<td>FGOALS-g2</td>
<td>3342</td>
<td>GFDL-CM4</td>
<td>3582</td>
</tr>
<tr>
<td>CanESM2</td>
<td>3582</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSIRO-Mk3.6.0</td>
<td>7236</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GFDL-ESM2M</td>
<td>7724</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. CMIP5 and CMIP6 models and their mean vertical resolution ($r_{vertical}$) in the stratosphere between 100 and 1 hPa (or the model top, whichever is lower).
As stated earlier, errors in the simulated \( f_{SSW} \) are associated with biases in the climatological mean strength and variability (on daily to interannual time scales) of the polar vortex. To assess to what extent the mean vortex strength is responsible, we recalculate \( f_{SSW} \) after replacing the \( U_{1060} \) daily climatology of each model with that of ERA-40 (Figs. 1c,d), and the remaining differences in \( f_{SSW} \) must be due to biases in vortex variability alone. Overall, \( f_{SSW} \) after the correction is in better agreement with ERA-40. In CMIP6, \( f_{SSW} \) improves in most cases, indicating that the models’ vortex variability is realistic and that biases in vortex strength are the main reason for their \( f_{SSW} \) errors. In contrast, even after the correction, many of the CMIP5 models still produce a too low \( f_{SSW} \). In fact, in around 57% of the CMIP5 models, \( f_{SSW} \) is still outside the observed range, and this must be caused by biases in vortex variability. Figure 1 also indicates that correcting the vortex strength improves the total \( f_{SSW} \), but this does not guarantee a realistic seasonal distribution of SSW. The correction leads to a substantial increase in late events for models with too-low overall frequencies (e.g., CCSM4, CESM2). On the other hand, after the correction, models with too high SSW frequencies (e.g., CanESM2, MPI-ESM1–2–HR) undergo a considerable reduction in December SSWs.

To summarize, CMIP5 and CMIP6 have somewhat different reasons for their biases in \( f_{SSW} \), motivating us to examine the two groups of models separately from each other. Doing so may also shed some new light on the recent progress in climate model development, as Fig. 1 already indicates that in terms of \( f_{SSW} \) CMIP6 performs better than CMIP5.

a. Variations in polar vortex strength and wave activity

As discussed in the introduction, we follow earlier work (Horan and Reichler 2017; Taguchi 2017; Martineau et al. 2018) and relate the climatological SSW frequency to the variability and strength of the polar vortex, and the vortex variability to the mean upward wave flux (Wu and Reichler 2019). In fact, Wu and Reichler (2019) demonstrated that the SSW frequency can be well estimated from the climatological strength of the polar vortex and upward wave activity fluxes. In the following analysis, we will pursue similar ideas by examining the vortex strength and wave activity to explain variations in SSW frequency.

Figures 2a and 2b show how variations in \( f_{SSW} \) relate to \( U_{1060} \). In this and the following figures, we consistently use color to represent individual models, with reddish for models with high \( f_{SSW} \) and blueish for models with low \( f_{SSW} \). As expected, the correlations between \( f_{SSW} \) and vortex strength are negative, with a tighter relationship for CMIP6 (\( r = -0.87 \)) than for CMIP5 (\( r = -0.62 \)). In CMIP6, \( U_{1060} \) alone is a very good indicator for \( f_{SSW} \). Many of its models are clustered around ERA-40 (Fig. 2b, black marker), indicating a reasonable performance. Also, the marker for ERA-40 is located at the regression line, another indicator for the consistency of CMIP6 with the reanalysis. In CMIP5, the linear fit between \( f_{SSW} \) and \( U_{1060} \) underestimates \( f_{SSW} \) of the “high-frequency” models (reddish markers) and also of the ERA-40. In the MMM (gray markers) and compared to ERA-40, the vortex of both ensembles is too strong and \( f_{SSW} \) is too low. The tendency of climate models to have overly strong polar vortices has been the subject of previous studies (e.g., Charlton et al. 2007), and Shaw et al. (2014) suspected that this could be due to a misrepresentation of the stratospheric gravity wave drag.

One may argue that variations in \( f_{SSW} \) impact the climatological polar vortex strength, so that the vortex strength cannot be considered as an independent estimator for \( f_{SSW} \). To address this concern, we repeat our analysis by excluding years with SSWs. In this case, the correlations are reduced but still sizeable (\( r = -0.44 \) in CMIP5, \( r = -0.77 \) in CMIP6) and significant at the 95% limit, suggesting that \( U_{1060} \) contains indeed distinct and independent information about the likelihood of SSWs. Therefore, we include in our subsequent analysis data from all years.

Figures 2c and 2d show, for each model, vertical profiles of the 40°–70°N averaged upward wave fluxes (FZ) relative to the ERA-40. In general, the intermodel variations increase with altitude, are larger in CMIP5 than in CMIP6, and are vertically quite uniform in CMIP6. The MMM of FZ (dashed black) is larger than ERA-40 (solid black) in the troposphere and smaller in the stratosphere, indicating that processes in the lower stratosphere limit the amount of wave activity that enters from below into the stratosphere. Of note is that stratospheric FZ is particularly underestimated in CMIP5. When comparing all panels, it becomes clear that models with low \( f_{SSW} \) (blueish and purplish colors) not only tend to be associated with a stronger \( U_{1060} \) but also with a reduced stratospheric wave activity compared to ERA-40.

b. Estimating the SSW frequency

In this section, we explore the usefulness of other climatological quantities beside \( U_{1060} \) as “predictors” for the intermodel variations in SSW frequency. As shown in Fig. 2b, in CMIP6, \( U_{1060} \) alone is already an excellent indicator for \( f_{SSW} \), making it difficult to make further improvements by including additional predictors in the linear regression.

Figure 3 shows for CMIP5 the actual versus the estimated \( f_{SSW} \) from linear regression, using the predictors indicated at the top of each panel. \( U_{1060} \) alone explains 38% of the total variance (Fig. 3a), with large errors at the two extremes of the distribution. The stratospheric wave driving (FZ) alone explains only 28% (Fig. 3b). We note that using FZ at levels higher than 100 hPa improves the explained variance (\( r^2 \)) (not shown) because of a closer relationship with the vortex at 10 hPa. However, we prefer FZ below 100 hPa as predictor because traditionally the 100-hPa level is considered as gateway for planetary waves entering the stratosphere, and the wave activity flux at 10 hPa is highly correlated with FZ.

Based on the ideas of Wu and Reichler (2019), we next combine the effects of \( U_{1060} \) and FZ by employing a bilinear regression to estimate \( f_{SSW} \) (Fig. 3c). The explained variance is now 70%, much improved over the single predictor regressions, and the regressions now also correctly predict \( f_{SSW} \) of the ERA-40. As expected, the regression coefficient of \( U_{1060} \) is negative and of FZ is positive (not shown), consistent with previous results (Jucker et al. 2014; Wu and Reichler 2019) that a weaker polar vortex and stronger stratospheric wave driving favor more SSWs.
The bar plots in Fig. 4a summarize our results by showing the explained variances \( r^2 \) from the linear regressions for CMIP5 and CMIP6. In a unary linear regression, the predictive power of both \( U_{1060} \) (blue) and \( FZ_{100} \) (orange) is stronger in CMIP6, and \( U_{1060} \) is the single best predictor for \( f_{SSW} \) in both ensembles. Combining \( U_{1060} \) and \( FZ_{100} \) (gray) leads to a large improvement in \( r^2 \) in CMIP5. The improvement in CMIP6 is more subtle, likely because \( r^2 \) from \( U_{1060} \) alone is already high. It is also of note that the intermodel variations \( U_{1060} \) and \( FZ_{100} \) in CMIP6 are negatively correlated \( (r = -0.47) \), while in CMIP5 this is not the case \( (r = 0.06) \). In other words, in CMIP6, \( U_{1060} \) and \( FZ_{100} \) are not independent, helping us to understand the only small improvements in explained variance when both predictors are used in combination. A negative relationship between \( U_{1060} \) and \( FZ_{100} \) is also revealed from their interannual variations in ERA-40 (not shown) and idealized models (Jucker et al. 2014; Wu and Reichler 2019). In this respect, the negative correlation between \( U_{1060} \) and \( FZ_{100} \) in CMIP6 is more plausible and realistic than the near-zero correlation in CMIP5. Repeating this analysis for only the high-top CMIP5 models or the CMIP5 models with a relatively fine vertical resolution does not change much the small correlations between \( U_{1060} \) and \( FZ_{100} \). This indicates that in CMIP5, reasons more complicated than the vertical structure of the model grid are responsible for the unphysical relationship between \( U_{1060} \) and \( FZ_{100} \).

We use bootstrapping to test the robustness of our regressions. Figure 4b shows the distributions of the regression coefficients \( \beta \), normalized by their own means, and derived from 50,000 regressions, in which we randomly determine the combination of models. Note that for the two bilinear regressions, we only present the coefficients of \( FZ_{100} \) (gray) and \( T_{TROPO} \) (green) (for definition see the caption of Fig. 3) since the ones of \( U_{1060} \) are in both cases highly significant. The regression coefficients are considered robust if their distributions are narrow, and they are significant if their 5th percentile is positive (i.e., of the same sign as the mean). As shown in Fig. 4b, this is the case for all regression models.
However, the fact that a bilinear regression coefficient passes the significance test does not imply that the second predictor adds real information. Therefore, we employ stepwise regression (Draper and Smith 1998, 307–312) to test whether the improvements of $r^2$ are significant at the 5% level when the second predictor is added. We find that for FZ100, this is the case in 95% (86%) of all 50,000 bootstrapping combinations of CMIP5 (CMIP6).

Given that $U_{1060}$ is the single best $f_{SSW}$ predictor, we next ask whether, besides FZ100, there are other more basic quantities that can be used as second predictors for the $f_{SSW}$ intermodel spread. Figure 5 shows latitude–pressure cross sections of the $r^2$ differences ($\Delta r^2$) between a bilinear regression using $U_{1060}$ and the quantity indicated at the top of each panel and a unary regression using $U_{1060}$ alone. Note that the actual $\Delta r^2$ is always positive, and that the color shading of $\Delta r^2$ in the plots indicates the sign of the second regression coefficient. In other words, reddish colors indicate a positive correlation and blueish colors indicate a negative correlation between $f_{SSW}$ and the second predictor. From Fig. 5, one can readily identify the locations where, in combination with $U_{1060}$, a second predictor improves explaining $f_{SSW}$.
Figure 5a shows $\Delta r^2$ for CMIP5 from using climatological zonal-mean zonal wind ($U$) as second predictor. Stronger $U$ in the midlatitudes favors the occurrence of SSWs, with the largest contributions in the UTLS region. Figure 5c indicates that in CMIP5, colder zonal-mean temperatures ($T$) in the UTLS region over the high latitudes also favor more SSWs. CMIP6 (right panels) shows similar relationships, even though the magnitude of $\Delta r^2$ is generally weaker. For example, colder $T$ along the high-latitude tropopause (Fig. 5d), stronger $U$ in the extratropics (Fig. 5b), and more upward wave fluxes (not shown) are all associated with higher $f_{\text{SSW}}$. Finally, Figs. 5e and 5f show the predictive capabilities of the index of refraction (RI$^2$), which conveniently explains the propagation of planetary waves and its dependence on the atmospheric background state. Planetary waves tend to propagate from regions of low RI$^2$ to regions of high RI$^2$ (Karoly and Hoskins 1982; Simpson et al. 2009). In both CMIP5 and CMIP6, there is a region centered at 40°N and 100 hPa where RI$^2$ is positively related to $f_{\text{SSW}}$.

**c. Role of the index of refraction**

To aid our subsequent analysis, we define from Fig. 5 three indices, each representing area averages of $U$, $T$, or RI$^2$ over specific regions: $U_{\text{TROPO}}$ is 35°–55°N and 200–70 hPa, $T_{\text{TROPO}}$ is 50°–90°N and 300–200 hPa, and RI$^2_{\text{TROPO}}$ is 35°–45°N and 200–70 hPa. These three indices are all located in the lower extratropical stratosphere, consistent with various studies that the conditions in this region are important for the upward propagation of planetary waves from the troposphere. For example, Gerber and Polvani (2009) suggested that strong winds and potential vorticity gradients in the lower stratosphere are essential for bursts of planetary waves into the vortex region, and thus for the creation of SSWs.

We examine the spatial structure of RI$^2$, focusing on stationary wave one, the dominant wave component of the stratosphere. Figures 6a and 6b show that the MMM of RI$^2$ in both ensembles is fairly similar, and it is also close to that of the ERA-40 (not shown). There is a region of maximum RI$^2$ in the middle to upper troposphere, which acts to trap the wave activity and corresponds to large EP flux convergences (Chen and Robinson 1992). More importantly, there is a band of small RI$^2$ just above the extratropical tropopause, which is a potential barrier to upward wave propagation and is associated with wave evanescence (Sigmond and Scinocca 2010). As shown by the black rectangles, the minima in RI$^2$ in both ensembles coincide more or less with our definition for RI$^2_{\text{TROPO}}$. This region also contains some negative values for CMIP5. The MMM of RI$^2_{\text{TROPO}}$ in CMIP6 is therefore somewhat larger than in CMIP5 (also see Figs. 7a,b). The cross sections of the intermodel spread in RI$^2$ (Figs. 6c,d) indicate that in both ensembles, the region of RI$^2_{\text{TROPO}}$ is characterized by large modeling uncertainties. Given that the expression for RI$^2$ [Eq. (2)] is dominated by various derivatives of the zonal-mean zonal wind, the large spread in RI$^2$ is indicative for difficulties in simulating the wind structure in this region. This is confirmed by Figs. 6e and 6f, showing the local correlations between intermodel variations in $U$ and RI$^2$. The correlations between RI$^2_{\text{TROPO}}$ and $U_{\text{TROPO}}$ are 0.8 in CMIP5 and 0.7 in CMIP6 (Table 2). In other words, over the RI$^2_{\text{TROPO}}$ region, RI$^2$ is very sensitive to uncertainties in the simulated wind, and from additional analysis (not shown) we find that it is mostly the $h_{52}$ term in (2) that explains the influence of $U$ on RI$^2$ in this region.

We next assess the role of RI$^2_{\text{TROPO}}$ in influencing the frequency of SSWs. Figures 7a and 7b show RI$^2_{\text{TROPO}}$ of the individual models plotted against their $f_{\text{SSW}}$. As expected, there is a positive relationship between RI$^2_{\text{TROPO}}$ and $f_{\text{SSW}}$, which is best explained from the increase in the upward-propagating wave activity with increasing RI$^2_{\text{TROPO}}$. Indeed, Table 2 shows that there is a positive correlation between RI$^2_{\text{TROPO}}$, $FZ_{100}$, and $FZ_{10}$. In terms of $T_{\text{TROPO}}$ (Figs. 7c,d), the MMM of both ensembles is ~3 K colder than ERA-40, but in terms of RI$^2_{\text{TROPO}}$ and $U_{\text{TROPO}}$ (not shown) both ensembles are about right. Also, $U_{\text{TROPO}}$ and $T_{\text{TROPO}}$ are only correlated at $r \sim 0.6$.

![Fig. 4. Outcome of different regression models to explain $f_{\text{SSW}}$. (a) Explained variance $r^2$. (b) Distribution of the normalized regression coefficients $\beta$, obtained from bootstrapping (see text for details); the coefficients are normalized by their own mean; box and whiskers indicate the 5th–95th-percentile range, minimum, maximum, and mean of the regression coefficients. For the two bilinear regressions, only the $\beta$ distributions for $FZ_{100}$ (gray) and $T_{\text{TROPO}}$ (green) are shown.](image-url)
indicating a certain degree of independence between the two quantities. We note that a cold lowermost stratospheric temperature bias is not exclusive to CMIP5 and CMIP6. For example, Polichtchouk et al. (2019) pointed out that all operational forecast systems at ECMWF suffer from such a bias over the tropical lower stratosphere, which depends on the vertical resolution and is associated with an inadequate representation of gravity waves. Figures 7c and 7d, and also Table 2 further show that $R_{TROPO}^2$ is negatively correlated with $T_{TROPO}$, a connection that may help explain why $T_{TROPO}$ has an explanatory power for $f_{SSW}$ (Figs. 5c,d). This issue will be explored next.

d. Relationship between tropopause temperature and SSW frequency

We begin by examining the predictive capabilities of $T_{TROPO}$ for $f_{SSW}$. Getting back to Fig. 3, using $T_{TROPO}$ on top of $U_{1060}$ as a second predictor (Fig. 3d) improves the estimation of $f_{SSW}$ in CMIP5, with $r^2$ increasing from 0.38 to 0.50. However, the
increase is smaller than from using FZ100 and U1060 (Figs. 3c and 4a). In CMIP6, using T\textsubscript{TROPO} as a second predictor also increases $r^2$, and this increase is similar to that from using FZ100 (Fig. 4a). The distributions of the regression coefficient $\beta$ for T\textsubscript{TROPO} are broader than for FZ100 (Fig. 4b), and in 61% (79%) of the bootstrapping cases T\textsubscript{TROPO} contains additional information about the $f_{SSW}$ estimates in CMIP5 (CMIP6). Therefore, FZ100 is a more robust second predictor than T\textsubscript{TROPO} in both ensembles. Nevertheless, we think that the connection between T\textsubscript{TROPO} and $f_{SSW}$ deserves further explanation.

We believe that T\textsubscript{TROPO} is connected to $f_{SSW}$ because a colder high-latitude tropopause is associated with a background state that favors the upward propagation of planetary waves into the polar vortex, leading to an increase in $f_{SSW}$. This
is consistent with previous studies, which also indicate that a colder lower stratosphere is linked to stronger $U$ and enhanced upward directed wave activity (Gerber and Polvani 2009; Schimanke et al. 2013; Shaw et al. 2014; Martineau et al. 2018). From the spatial correlations between $T_{TROPO}$ and $U_{TROPO}$ we find a band of strong negative correlations that tilts upward and northward from the tropopause at 40°N into the stratosphere (not shown), as one would expect from the thermal wind relationship. This band is similar in structure to the improvements in $r^2$ from using $U$ as a second predictor for $f_{SSW}$ (Figs. 5a,b), suggesting that the improvements are related to variations in $T_{TROPO}$. As given in Table 2, the correlations between $T_{TROPO}$ and $U_{TROPO}$ are $-0.6$ in both ensembles because of the thermal wind relationship, leading to similar correlations between $T_{TROPO}/U_{TROPO}$ and $R_{TROPO}^2$. The correlation between $T_{TROPO}$ and $f_{SSW}$ is as expected negative, but it is also quite weak ($r = 0.2–0.3$; Table 2), reflecting the complicated chain of nonlinear dynamical processes that link $T_{TROPO}$ to $f_{SSW}$, and which involve, among others, $R_{TROPO}^2$ and $FZ_{100}$.

We want to point out that our analysis of Table 2 and Fig. 7 is solely based on correlations, which do not establish causality. To partly address the question of cause and effect between $T_{TROPO}/U_{TROPO}$ and $FZ$, we examine lagged correlations from daily ERA-40 (Fig. 8). The correlations are generally small ($r = -0.2$), but given that these are daily data for 44 years, the correlations are significant. The correlations are strongest when $T_{TROPO}$ leads by about one week (Fig. 8a), suggesting

![Fig. 7. Intermodel variations. (a),(b) $R_{TROPO}^2$ vs $f_{SSW}$; $R_{TROPO}^2$ is normalized by the value from ERA-40. (c),(d) $T_{TROPO}$ vs $R_{TROPO}^2$. Colors show individual models, as in Fig. 2; black represents the ERA-40 and gray the MMM. Regression lines are also shown, with correlation coefficients given in the plots.](http://journals.ametsoc.org/jcli/article-pdf/33/23/10305/5014458/jclid200104.pdf)

### Table 2. Correlation of the intermodel spread in various indices. For each index, the top number is for CMIP5, and the bottom number is for CMIP6. Correlations followed by an asterisk (*) are based on absolute differences in $T_{TROPO}$, $U_{TROPO}$, and $R_{TROPO}^2$ with respect to their MMM. Correlations significant at the 95% level are shown in boldface and at the 90% level in italic.

<table>
<thead>
<tr>
<th>$T_{TROPO}$</th>
<th>$U_{TROPO}$</th>
<th>$R_{TROPO}^2$</th>
<th>$FZ_{400}$</th>
<th>$FZ_{40}$</th>
<th>$f_{SSW}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{TROPO}$</td>
<td>-0.6</td>
<td>-0.5</td>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.2</td>
</tr>
<tr>
<td>$U_{TROPO}$</td>
<td>0.8</td>
<td>0.7</td>
<td>0.4</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>$R_{TROPO}^2$</td>
<td>0.7</td>
<td>0.4</td>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_{vertical}$</td>
<td>0.3*</td>
<td>0.1*</td>
<td>0.1*</td>
<td>-0.6</td>
<td>-0.6</td>
</tr>
<tr>
<td>$0.6*$</td>
<td>0.3*</td>
<td>0.1*</td>
<td>-0.5</td>
<td>-0.7</td>
<td>-0.4</td>
</tr>
</tbody>
</table>
that variations in $T_{TROPO}$/$U_{TROPO}$ from synoptic variability indeed influence the upward propagation of planetary waves into the stratosphere. On the other hand, the correlations between FZ$_{100}$ and $T_{10}$, the 10-hPa temperatures averaged from 50°-90°N, peak when FZ$_{100}$ leads (Fig. 8b), indicating that midstratospheric warming is a consequence of upward-propagating waves and their poleward heat transports.

e. The role of vertical model resolution

The final question we attempt to address is whether perhaps a model’s vertical resolution is connected to the simulation of stratospheric variability, similar to the finding that the location of a model’s top is important (Shaw and Perlwitz 2010; Charlton-Perez et al. 2013). First, we investigate potential causes for the biases in $T_{TROPO}$. Such biases must be accompanied by variations in the height of the thermally defined tropopause, as has been previously reported for CMIP5 (Kim et al. 2013; Ao et al. 2015). The mean tropopause pressure (Figs. 9a,b) has a larger intermodel spread in CMIP5 than in CMIP6, but the MMM of it (dashed black) is similar between the two ensembles and low-biased with respect to the reanalysis (solid black). The upward shifted tropopause corresponds to a cold bias above and/or a warm bias below the tropopause. Similar cold biases and too high tropopause heights have been reported before (SPARC 2010; Charlton-Perez et al. 2013), indicating that these are common climate model problems.

As mentioned in the introduction, sufficient vertical resolution is an important prerequisite for properly representing the UTLS region, the location of the tropopause, and the propagation of waves. We use the vertical resolution in the stratosphere ($r_{vertical}$) to examine potential influences of it on the simulation of $T_{TROPO}$. Figures 9c and 9d show the correlation between $r_{vertical}$ and the absolute temperature difference ($\Delta T$) with respect to the MMM at each grid point. Here, we use absolute differences because a coarse vertical resolution can lead to either upward or downward displacements of the tropopause. The region of reddish colors above the extratropical tropopause (~250 hPa) indicates that a coarse resolution (larger $r_{vertical}$) is indeed associated with larger errors in $T_{TROPO}$. However, the correlation of $r_{vertical}$ with $U_{TROPO}$ and $R_{TROPO}$ is much smaller (Table 2), perhaps because of the complicated nonlinear nature of $R_{TROPO}$. Repeating this analysis using the vertical resolution in the UTLS region leads to very similar results (not shown).

Last, we examine how $r_{vertical}$ is connected to the simulated stratospheric variability. Table 2 demonstrates that coarse vertical resolution is associated with reduced upward wave propagation (FZ$_{100}$ and FZ$_{10}$) and $f_{SSW}$. However, $r_{vertical}$ is only weakly correlated with $|\Delta U_{TROPO}|$ and $|\Delta R_{TROPO}|$, suggesting that the associations between $r_{vertical}$ and FZ$_{100}$, FZ$_{10}$, and $f_{SSW}$ are more related to the improved representation of stratospheric wave dynamics when the vertical resolution is high, and not so much to the influence of resolution on biases in $T_{TROPO}$ and $U_{TROPO}$.

4. Summary, discussion, and conclusions

We investigate the intermodel variations in the SSW frequency simulated by the two latest generations of climate models, CMIP5 and CMIP6. Both generations of models show a wide spread in their frequencies, with some models producing no SSW, and others having more than one SSW per year. This raises concerns about the models’ ability to simulate a realistic stratospheric circulation variability, the downward influence of this variability into the troposphere, and potential shift of this variability under future climate change. The main goal of this study is to identify basic indicators that can be used to understand why so many models produce unrealistic SSW frequencies and how to improve the models. To this end, we compare the intermodel spread of various quantities. We find that the climatological mean strength of the polar vortex is the single most reliable indicator for the SSW frequency, followed by the upward-propagating wave activity flux in the lower stratosphere. In many models, this flux undergoes an unrealistic reduction in the lower stratosphere, which is connected to biases...
in the simulation of the atmospheric background in this region and also to a too coarse vertical model resolution. In agreement with earlier studies (Scott and Polvani 2004; Son et al. 2007), we identify a critical region in the lower stratosphere just above the subtropical jet, where the index of refraction for planetary waves has a local minimum and acts as a valve controlling the vertical wave propagation into the stratosphere. This is indicated by a robust correlation between the upward wave activity flux and the index of refraction in this region. The index is highly sensitive to subtle inaccuracies in the simulation of the zonal-mean zonal wind, which in turn are connected to temperature biases above the high-latitude tropopause through the thermal wind relationship. Overall, our results are also consistent with findings with idealized models (Jucker 2016; Martineau et al. 2018). However, it is unclear to what extent a model’s high-latitude temperatures actively influence the SSWs through the thermal wind relation, or whether it is primarily the zonal-mean zonal wind that controls the SSWs and the temperatures over the high latitudes simply follow along.

Our results for CMIP5 and CMIP6 are often similar, for example in terms of the sign and magnitude of the relationships and the location of key regions. This similarity adds to the credibility of our findings. Since we perform our analysis separately for CMIP5 and CMIP6, we are also able to judge the overall performance of the two ensembles. We find that CMIP6 as a whole is more realistic than CMIP5. This is most obvious from the smaller rms error in the simulated SSW frequency (26% vs 37%), but various correlations in CMIP6 are also more robust and physical than in CMIP5. This indicates that the stratospheric simulation performance in CMIP5 is influenced by a more diverse range of errors than in CMIP6, making it more difficult to linearly relate errors to specific sources. We believe that besides enhanced parameterization, finer resolutions and higher model tops are important reasons for the improvements seen in CMIP6: the mean of all CMIP6 models has a top at 1 hPa and a vertical resolution in the stratosphere (UTLS region) of ~1.8 km (~900 m), whereas in CMIP5 the mean top is located at 2.6 hPa and the vertical resolution is ~2.9 km (~1300 m). The finer vertical resolution in CMIP6 creates a more realistic lower stratospheric background, improves the simulation of the stratospheric wave dynamics, and reduces the negative biases in the vertical wave activity flux, and hence in SSW frequency. In an earlier study, Charlton-Perez et al. (2013) argued that models with a low model lid
(below 1 hPa) tend to reproduce too few SSWs, and that a high lid is needed for a model to produce a realistic frequency. Indeed, we find that the SSW frequency and the height of the model lid are correlated at \( r \approx 0.6 \) in CMIP5. In CMIP6, however, this correlation reduces to \( r \approx 0.3 \), which is smaller than \( r \approx -0.4 \) between the vertical resolution and the SSW frequency (Table 2). There is also some linear relationship between the vertical resolution and the height of the model lid \( (r \approx -0.6 \text{ in both ensembles}) \), making it difficult to determine whether the improvements in the SSW frequency of CMIP6 are due to the finer resolution, a higher lid, or a combination of both.

There are also some limitations to our study. For example, our results depend on a mix of CMIP models. These models form ensembles of opportunity and do not span the range of uncertainties desirable for this study, creating unavoidable imperfections. Another potential limitation to our study is the use of all available simulation years to calculate the needed climatologies and estimate the frequency of SSWs. We did so because no prior knowledge about a model’s SSWs is needed. Some climatologies, like the strength of the stratospheric winds, are influenced by SSWs and are therefore not independent estimators. An alternative approach would have been to only include years without SSWs, but we found that doing so does not fundamentally change our conclusions. Our study is also limited in the sense that the reasons for the intermodel spread in polar vortex strength remain unclear, despite our finding that this strength is crucial for a model’s SSW frequency. However, biases in polar vortex strength may be related to a near endless list of reasons, and in most cases it is impossible to diagnose these reasons from the available model output. One such reason may be the treatment of gravity waves. For example, Fig. 1 shows that there are surprisingly large differences in the SSW frequencies produced by the Canadian CanESM2 (110%) and CanESM5 (40%), despite the structural similarities of these two models. These differences are likely related to changes in the orographic gravity wave drag parameterizations (J. Anstey 2019, personal communication). Data for the gravity wave drag tendencies, which could be used to clarify the role of gravity wave drag for SSWs, are available from the CMIP6 archives, but this task must be left to future research. Another2 lesson learned from this study is that it is difficult to pinpoint errors in a model’s SSW frequency to other simulation biases, simply because the frequency depends on the subtle interplay of a variety of complicated factors. Simulating a reasonable SSW frequency can therefore be considered as a sensitive indicator for a model’s overall stratospheric simulation performance.

A potential application of the results of this study is that a model’s expected SSW frequency can be estimated from only two basic modeling aspects: vortex strength and lower stratospheric wave activity flux. A model developer can derive both quantities at reasonable accuracy from relatively short simulations, avoiding the necessity of much longer simulations for more robust SSW statistics.

Acknowledgments. We thank the National Science Foundation under Grant 1446292 and the Department of Atmospheric Sciences at the University of Utah for their support. The use of the computing infrastructure from the Center for High Performance Computing at the University of Utah is gratefully acknowledged. We thank Isla Simpson for valuable suggestions and comments, and Andrew Charlton-Perez and the other two anonymous reviewers for their comments that helped to improve this paper. We acknowledge ECMWF for providing the ERA-40 data, and the Coupled Model Intercomparison Project, the World Climate Research Programme, and the various modeling groups for producing and making available the model output used in this study.

REFERENCES


