A New Look at Stratospheric Sudden Warmings. Part II: Evaluation of Numerical Model Simulations

ANDREW J. CHARLTON,[*],[@@] LORENZO M. POLVANI,[+] JUDITH PERLWITZ,[#] FABRIZIO SASSI,[@]
ELISA MANZINI,[&] KIYOTAKA SHIBATA,[**] STEVEN PAWSON,[++] J. ERIC NIELSEN,[++] AND DAVID RIND[##]

[*] *Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York*
[+] *Department of Applied Physics and Applied Mathematics, and Department of Earth and Environmental Sciences, Columbia University, New York, New York*
[#] *Cooperative Institute for Research in Environmental Sciences, Climate Diagnostics Center, University of Colorado, and Physical Sciences Division, NOAA/Earth System Research Laboratory, Boulder, Colorado*
[@] *National Center for Atmospheric Research, Boulder, Colorado*
[&] *Istituto Nazionale di Geofisica e Vulcanologia, Bologna, Italy*
[**] *Meteorological Research Institute, Tsukuba, Ibaraki, Japan*
[++] *Global Modeling and Assimilation Office, NASA GSFC, Greenbelt, Maryland*
[##] *NASA Goddard Institute for Space Studies, New York, New York*

(Manuscript received 13 October 2005, in final form 28 March 2006)

ABSTRACT

The simulation of major midwinter stratospheric sudden warmings (SSWs) in six stratosphere-resolving general circulation models (GCMs) is examined. The GCMs are compared to a new climatology of SSWs, based on the dynamical characteristics of the events. First, the number, type, and temporal distribution of SSW events are evaluated. Most of the models show a lower frequency of SSW events than the climatology, which has a mean frequency of 6.0 SSWs per decade. Statistical tests show that three of the six models produce significantly fewer SSWs than the climatology, between 1.0 and 2.6 SSWs per decade. Second, four process-based diagnostics are calculated for all of the SSW events in each model. It is found that SSWs in the GCMs compare favorably with dynamical benchmarks for SSW established in the first part of the study.

These results indicate that GCMs are capable of quite accurately simulating the dynamics required to produce SSWs, but with lower frequency than the climatology. Further dynamical diagnostics hint that, in at least one case, this is due to a lack of meridional heat flux in the lower stratosphere. Even though the SSWs simulated by most GCMs are dynamically realistic when compared to the NCEP–NCAR reanalysis, the reasons for the relative paucity of SSWs in GCMs remains an important and open question.

1. Introduction

In Part I of this study (Charlton and Polvani 2006, henceforth CP06), we constructed a new climatology of major midwinter stratospheric sudden warmings (SSWs) and proposed benchmarks for their simulation in general circulation models (GCMs). In this study we will analyze the simulation of SSWs by a series of stratosphere-resolving GCMs. The GCMs will be evaluated in two ways. First, the number, type, and climatology of SSWs in the models will be compared to the climatology established by CP06. Second, process-based benchmarks of SSWs, introduced by CP06, will be used to assess the performance of each GCM.

Previous studies have examined the simulation of SSWs by individual stratosphere-resolving GCMs (e.g., Butchart et al. 2000; Manzini and Bengtsson 1996; Erlebach et al. 1995), but as far as we are aware there has been no comprehensive intercomparison of the performance of a series of GCMs in this respect. Most of the recent intercomparisons of stratosphere-resolving GCMs (e.g., Austin et al. 2003; Shine et al. 2003) have touched only briefly on the simulation of SSWs.

The occurrence of SSWs is crucial to the chemistry of ozone, since the low temperatures that occur in undisturbed winters are an important prerequisite for deni-

[@@] Current affiliation: Department of Meteorology, University of Reading, Reading, United Kingdom.

Corresponding author address: Andrew J. Charlton, Department of Meteorology, University of Reading, Reading, Berkshire, RG6 6BB, United Kingdom.
E-mail: a.j.charlton@reading.ac.uk

trification and subsequent catalytic ozone loss, if the vortex remains intact into the spring. The importance of warmings was recognized early in the GCM Reality Intercomparison Project for SPARC (GRIPS) model evaluation (Pawson et al. 2000), but the restricted length of model simulations available until quite recently has precluded detailed examination of the frequency of occurrence of simulated SSWs. With longer runs of coupled Chemistry–Climate Models (CCMs) now possible, the CCM Validation (CCMVal) project (Eyring et al. 2005) will examine SSWs in more detail. This study, along with CP06, should complement and inform CCMVal, both by assessing the performance of some GCMs and by suggesting process-based dynamical benchmarks to test other GCMs.

Part of the interest in validating stratosphere-resolving GCMs is also due to the potential interactions between greenhouse gas–induced climate changes, stratospheric ozone depletion, and dynamical coupling between the stratosphere and troposphere (Hartmann et al. 2000). There is little consensus about future changes to variability of the Arctic stratospheric polar vortex (Rind et al. 1998; Schnadt and Dameris 2003). We surmise that a necessary though not sufficient condition for the suitability of GCMs to accurately simulate future stratospheric variability is that they produce a credible simulation of the current SSW climatology. Other factors, such as the simulation of future tropospheric variability, should also be considered when determining the suitability of a GCM for this task.

The paper is structured as follows. The GCMs to be analyzed and the methods used are described in section 2. In section 3 we compare the stratospheric climatology of the GCMs. In section 4 we examine the number, type, and climatology of SSWs. In section 5 we compare process-based benchmarks of SSWs between the GCMs. In section 6 we provide further discussion and comparison of the stratospheric dynamics of each GCM. In section 7 we present conclusions.

2. Methodology and GCM runs

This section briefly describes the methodology used to identify and classify SSWs and gives brief details of the GCMs used in the study. Neither discussion is intended to be exhaustive and readers should consult relevant references for further details.

The methodology for identifying and classifying SSWs is described in full by CP06. We confine our study to SSWs that occur during the extended winter season, November to March. First, SSWs are defined to occur when the zonal mean zonal wind at 10 hPa and 60°N becomes easterly, in line with the WMO definition. An additional criterion, that the zonal mean zonal winds return to westerlies for 10 or more consecutive days following the SSW, is used to remove events that are final warmings. Second, the algorithm classifies SSWs into vortex splits (in which the stratospheric polar vortex breaks into two comparably sized pieces) and vortex displacements (in which the vortex remains largely intact). The algorithm uses absolute vorticity to identify the vortex edge and then compares the size and strength of cyclonically rotating vortices in the flow to determine if events are vortex splits or vortex displacements.

In CP06, data from both the National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP–NCAR) reanalysis (Kistler et al. 2001), and its Climate Data Assimilation System (CDAS) extension, and the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40) dataset (Kallberg et al. 2004) were used to establish a climatology of SSWs events between the winter seasons of 1957/58 and 2001/02. The results from the two reanalysis datasets were found to be very similar, as should be expected given their largely common source of observations. Therefore in the present study we use only the NCEP–NCAR data to evaluate the GCMs. This also makes the construction of many of the statistical tests much simpler.

The CP06 algorithm was used to test the simulation of SSWs in a series of GCM simulations. We study GCMs that are explicitly designed to resolve the stratosphere, which we call stratosphere resolving. We define stratosphere-resolving GCMs as those with a model top close to or above the stratopause (approximately 50 km or 0.8 hPa) and with a meaningful number of model levels (10 or more) in the stratosphere. One major constraint in choosing and obtaining GCM integrations in order to examine the intra-annual variability of SSWs is that daily or finer time resolution of diagnostic fields is required. We found that the archiving of daily output is by no means a standard practice among the modeling centers and groups that run stratosphere-resolving GCMs.

The GCMs used in this study are summarized in Table 1, and the forcings used in each model are shown in Table 2. In the following subsections we briefly discuss each GCM. We have attempted to restrict our attention in this study to GCM runs that are forced by sea surface temperatures (SSTs) from the same time period as the NCEP–NCAR reanalysis. It was not possible to obtain runs with observed SSTs for the Meteorological Research Institute/Japan Meteorological Agency 1998 Model (MRIJMA; run with climatological SSTs), and this may be a potential source of bias.

We have also attempted to examine the longest avail-

TABLE 1. GCM experiments used in the study.

GCM	Run length/winters	SST forcing	Horizontal resolution	Vertical levels	Model top	Reference
FVGCM	49	Obs 1949–97	2° × 2.5°	55	0.01 hPa	Stolarski et al. (2005)
GISSL53	47	Obs 1951–97	4° × 5°	53	0.002 hPa	Rind et al. (2002)
GISSL23	46	Obs 1951–96	8° × l0°	23	0.002 hPa	Shindell et al. (1998)
WACCM	50	Obs 1951–2000	T63	66	150 km	Sassi et al. (2004)
MAECHAM	29	Obs 1970–98	T42	39	0.01 hPa	Manzini et al. (2006)
MRIJMA	60	Climate 60 years	T42	45	0.01 hPa	Shibata et al. (1999)

able runs of each GCM, to try to avoid spurious disagreement between the GCMs and reanalysis resulting from potential decadal variability of SSWs (Butchart et al. 2000). Except for the Middle Atmosphere ECHAM Model (MAECHAM; which is run for 29 full winter seasons), all of the GCM runs used here are over comparable or longer time periods than the reanalysis data, typically 50 yr.

a. NASA Goddard Space Flight Center, finite-volume GCM (FVGCM)

The National Aeronautics and Space Administration (NASA) Goddard Earth Observing System (GEOS-4) GCM is a middle-atmosphere GCM based on the finite-volume dynamical core of Lin (2004), with gravity wave drag, cloud, and cumulus parameterizations originally based on those in the Community Atmosphere Model version 3 (CAM3). The model has 55 levels in the vertical and the model top is at 0.01 hPa (approximately 80 km); the average vertical spacing of levels in the stratosphere is 1.2 km. The model has a flexible horizontal resolution and is run in this case at 2° × 2.5°. The model is forced with observed SSTs and sea ice between 1949 and 1997 using the Rayner et al. (2003) dataset. The model runs are described in more detail in Stolarski et al. (2005).

b. NASA Goddard Institute for Space Studies, Global Climate/Middle Atmosphere Model, new version (GISSL53)

The new NASA Goddard Institute for Space Studies, Global Climate/Middle Atmosphere Model 3 is an update from the previous version of the Global Climate–Middle Atmosphere Model (GCMAM; Rind et al. 2002). The update includes new boundary layer and turbulent schemes, convective and cloud cover parameterizations, and atmospheric radiation code. The broad nature of the changes to these schemes is shared with the new GISS model-E (Schmidt et al. 2006). The gravity wave drag in this model utilizes the formulations discussed in Rind et al. (1999, 1988) except that much smaller values are used. A major difference with the other models is that the nonorographic gravity wave drag components are a function of resolved processes in the troposphere. The model has four different vertical and horizontal resolutions; the version used here is 4° × 5°, with 53 layers in the vertical and model top at 0.002 hPa (approximately 85 km). The vertical spacing is 500 m in the middle to upper troposphere, 0.5 to 1 km in the lower stratosphere, and 2 to 2.5 km in the upper stratosphere. The model is forced with observed SSTs and sea ice between 1951 and 1997 using the Rayner et al. (2003) dataset. A more complete description of all the versions of the model is given in Rind et al. (2006, manuscript submitted to *J. Geophys. Res.*).

c. NASA Goddard Institute for Space Studies, Global Climate/Middle Atmosphere Model, legacy version (GISSL23)

The NASA GISS Global Climate/Middle Atmosphere Model is a middle-atmosphere GCM based on the climate model of Hansen et al. (1983) and the middle-atmosphere version outlined by Rind et al.

TABLE 2. Forcings used in each GCM run.

GCM	Sea ice extent	CO_2 conc./ppmv	O_3 climatology	Solar forcing TOA/W m^{-2}
FVGCM	Obs 1949–97 (Rayner et al. 2003)	Fixed 355	Monthly variance (Langematz 2000)	Fixed 1367
GISSL53	Climate 1975–84 (Rayner et al. 2003)	Fixed 311	Monthly and yearly variance; multiple sources (see text)	Fixed 1365.5
GISSL23	Climate 1975–84 (Rayner et al. 2003)	Fixed 311	Monthly variance (London et al. 1976)	Fixed 1367.6
WACCM	Obs 1951–2000 (NCEP–NCAR, Reynolds)	Fixed 355	Monthly variance (Liang et al. 1997)	Fixed 1367
MAECHAM	Obs 1970–98 (Rayner et al. 2003)	Fixed 348	Monthly variance (Fortuin and Kelder 1998)	Fixed 1365
MRIJMA	Climate 1978–98 (Rayner et al. 2003)	Fixed 348	Monthly variance (Liang et al. 1997) (<0.4 hPa) CIRA (>0.4 hPa)	Fixed 1365

(1988). The model has 23 levels in the vertical and the model top is at 0.002 hPa (approximately 85 km). The vertical spacing of levels is 0.2 km near the surface, 3.8 km in the upper troposphere, and 5 to 5.8 km in the stratosphere. The model has a horizontal resolution of 8° × 10°. The model is forced with observed SSTs and sea ice between 1951 and 1997 using the Rayner et al. (2003) dataset. GISSL23 has a much coarser horizontal and vertical resolution than most of the other models in the study. We include it because it has been used in a number of high-profile studies that examined the response of the stratosphere to changing greenhouse gas concentrations and the impact of these changes on the tropospheric flow (e.g., Shindell et al. 1999). Note that this version of the model differs from that used in Rind et al. (1988) and subsequent publications in that it has greatly reduced orographic drag (Shindell et al. 1998).

d. NCAR Whole Atmosphere Community Climate Model (WACCM)

The NCAR Whole Atmosphere Community Climate Model version 1b is an extended version of the NCAR Community Climate Model version 3 (CCM3; Kiehl et al. 1998). The model has 66 levels in the vertical and the model top is at 150 km (approximately 0.000002 hPa). The average vertical spacing of levels in the stratosphere is 1.5 km. The model has a spectral formulation, with resolution of T63 (approximately 1.875° × 1.875°). The model is forced with observed SSTs from 1950 to 2000 using the NCEP Reynolds observed dataset (http://podaac.jpl.nasa.gov/reynolds). The model runs are described in more detail in Sassi et al. (2004).

e. Max Planck Institute for Meteorology (MPI)/Middle Atmosphere ECHAM Model (MAECHAM)

The MPI MAECHAM model is an extended version of the MPI ECHAM5 model (Roeckner et al. 2003). The model has 39 levels in the vertical and the model top is at 0.01 hPa (approximately 80 km). The vertical spacing of levels in the stratosphere varies from 1.5 to 3 km. The model has a spectral formulation, with resolution of T42 (approximately 2.8° × 2.8°). The model is forced with observed SST and sea ice forcings, from the Atmospheric Model Intercomparison Project II (AMIP II; Gates et al. 1999). The model is described in more detail by Manzini et al. (2006).

f. The Meteorological Research Institute/Japanese Meteorological Agency 1998 Model (MRIJMA)

The MRIJMA 1998 Model is a hybrid version of the Meteorological Research Institute model (Chiba et al. 1996) and the operational global model (GSM9603) of the Japan Meteorological Agency (JMA 1997). The model has 45 levels in the vertical and the model top is at 0.01 hPa (approximately 80 km). The average vertical spacing of levels in the stratosphere is 2 km. The model has a spectral formulation, with resolution of T42 (approximately 2.8° × 2.8°). The model is forced with climatological SSTs and run for 60 yr. The model setup is described in more detail in Shibata et al. (1999). The climatological SSTs are 21-yr averages between 1978 and 1998, based on the Hadley Centre SST dataset (HadSST).

3. Climatology of GCMs

In this section, the stratospheric climatology of the GCMs is briefly examined. An indication of the strength and size of the stratospheric polar vortex in each GCM can be gained by examining the stratospheric climatology, with the caveat that it is often difficult to separate the time and zonal mean state of the stratosphere and its time-varying component. We restrict our analysis to the zonal mean zonal wind at 10 hPa and the meridional heat flux at 100 hPa both for sake of brevity and because of the limited amount of data available to us.

a. Zonal mean zonal wind at 10 hPa

The zonal mean zonal wind on the 10-hPa pressure surface as a function of latitude and time for each of the GCMs and the NCEP–NCAR reanalysis is shown in Fig. 1. The top-middle and top-right panels show line plots of the winter mean zonal mean zonal wind as a function of latitude for the various GCMs (colored lines) and for the NCEP–NCAR reanalysis (black line). There is large variability between the GCMs, both in the seasonality of the zonal mean zonal wind and in its maximum and winter mean values.

In terms of the winter mean zonal mean zonal wind, three GCMs have a zonal jet either within or close to one standard deviation from the NCEP–NCAR reanalysis. Only GISSL23 and WACCM have zonal wind speeds noticeably different from the reanalysis. Both have very strong winter mean zonal mean wind maximum—GISSL23 has a maximum of 43.4 m s^{-1} while WACCM has a maximum of 44.8 m s^{-1} compared to the NCEP–NCAR reanalysis maximum of 21.9 m s^{-1}. The extremely strong jets in these GCMs are reminiscent of the “cold pole” problem prevalent in many stratosphere resolving GCMs (Pawson et al. 2000). A further curiosity is the easterly zonal mean zonal wind values close to the pole in GISSL23. The extremely strong jets in the GISSL23 model are a direct result of the reduced orographic drag used by Shindell et al.

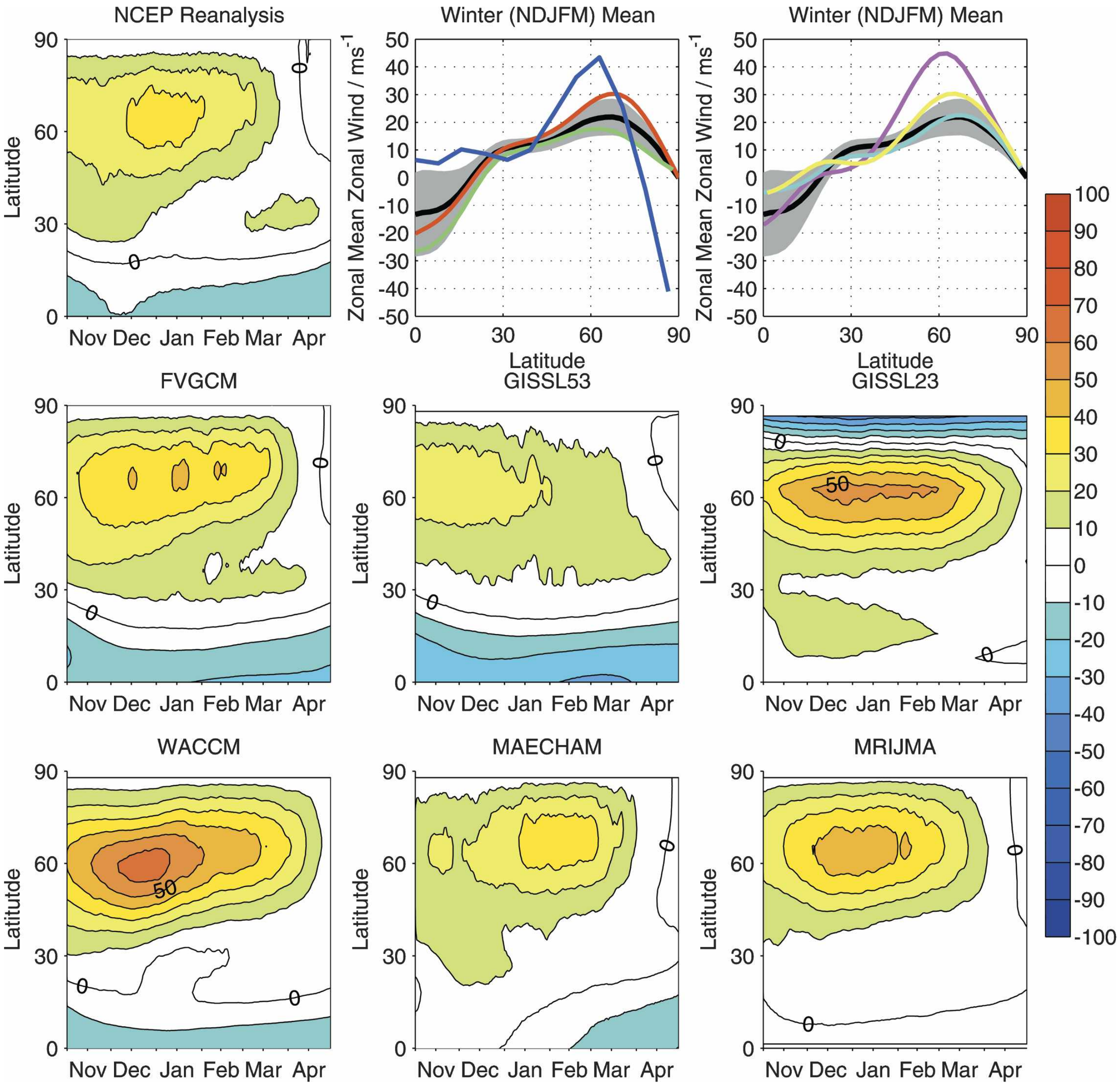

FIG. 1. Zonal mean zonal wind climatology at 10 hPa for GCMs that resolve the stratosphere in this study. (top left) Climatology from NCEP–NCAR reanalysis for the years 1958–2002. Contour interval is 5 m s^{-1}. (top middle), (top right) Line plots show winter mean for each GCM: NCEP–NCAR climatology in thick black line, FVGCM in red line, GISSL53 in green line, GISSL23 in blue line, WACCM in magenta line, MAECHAM in the cyan line, and MRIJMA in yellow line. Gray shading shows ±one interannual standard deviation from the mean.

(1998) and are not a characteristic of the model as normally used (e.g., Rind et al. 1988, their Figs. 2 and 3). Only GISSL53 has a weaker winter mean zonal mean wind maximum (17.6 m s^{-1}) than the reanalysis.

The seasonal cycle [Fig. 1 and further analysis (not shown)] varies markedly between the different GCMs. The reanalysis shows peak zonal mean zonal winds in the extratropics between days 50 and 80 of the winter season (late December to early February). Three of the models (FVGCM, WACCM, and MRIJMA) simulate this seasonality correctly, although the absolute values of the zonal mean zonal wind in WACCM are on average 15 m s^{-1} larger than the NCEP–NCAR reanalysis. GISSL53 has a seasonality shifted toward early winter, with peak zonal mean zonal winds between days 10 and 30 (November). MAECHAM has a seasonality shifted toward late winter, with peak zonal mean zonal winds between days 80 and 110 (late January to February). GISSL23 has a much broader zonal mean zonal wind peak than the reanalysis or the other models, with large relative zonal wind speeds throughout February. Most of the models simulate the climatology of March and April well, although GISSL53 has absolute values of zonal mean wind speed at 60°N and 10 hPa slightly smaller than the reanalysis in early March. It is also noticeable that in both GISS23 and WACCM the final warming is much later than in other GCMs or the NCEP–NCAR reanalysis. Also note that none of the models produces a recognizable quasi-biennial oscillation (QBO).

b. *Meridional heat flux at 100 hPa*

Throughout the rest of this study, the meridional heat flux ($\overline{v'T'}$) will be used as a proxy for the vertical component of the Eliassen–Palm flux, following Polvani and Waugh (2004). We use the meridional heat flux, rather than the full Eliassen–Palm flux (Andrews et al. 1985), as we were only able to obtain limited amounts of data on daily time scales. In particular, only information on 100- and 10-hPa pressure surfaces was obtained, making it impossible to calculate vertical derivatives, which are required to calculate the full Eliassen–Palm flux.

The meridional heat flux climatology (Fig. 2) at 100 hPa is noisy, even in the NCEP–NCAR reanalysis plot, highlighting its large interannual variability. The NCEP–NCAR meridional heat flux has a very broad seasonality, with large values occurring between mid-November and early March. Peak values of heat flux are centered at 60°N throughout winter. The meridional heat flux at 100 hPa is well simulated by most of the GCMs. Apart from GISSL23, the seasonality is well simulated. WACCM and MRIJMA have a heat flux climatology peaked too strongly in midwinter with large values in December and January, but their wintertime means are very close to the NCEP–NCAR climatology.

Only the two GISS GCMs show major differences with the NCEP–NCAR climatology. GISSL53 has a very broad band of positive heat flux and a peak located approximately 10 degrees of latitude south of the peak in the NCEP–NCAR reanalysis. GISSL23 has very weak heat flux throughout the year, and a wintertime mean peak heat flux less than 50% that of the NCEP–NCAR reanalysis. Part of this discrepancy may be due to the relatively coarser horizontal resolution in the GISS models.

4. Stratospheric warmings in the GCMs

In this section GCMs are evaluated by comparing the statistics of SSWs with the statistics obtained from the NCEP–NCAR reanalysis, which can be found in Table 2 of CP06.

a. *Frequency of major midwinter warmings*

The number of SSWs in the GCM simulations is shown in Table 3. We compare the mean frequency of SSWs per winter to take into account the different length of each GCM run and the reanalysis dataset. Column 2 shows the length of the model run, column 3 shows the number of SSWs recorded, and column 4 shows the expected frequency of SSWs per winter. The standard error of the frequency estimate is shown in column 5. A *t* test is used to determine which models have a significantly different frequency of SSWs compared to the NCEP–NCAR reanalysis dataset. For more details of the statistical procedure see appendix A. Columns 5 and 6 show if the frequency of SSWs in each model is significantly different from the frequency of SSWs in the NCEP–NCAR reanalysis at significance levels of 0.10 and 0.05.

Most of the GCMs in the present study have a lower frequency of SSWs than the reanalysis datasets. Of the six GCM simulations, only one, GISSL53, has a higher frequency of SSWs than the reanalysis datasets. Three of the five GCMs that underestimate the frequency of SSWs are significantly deficient at both the 0.10 and 0.05 confidence levels (GISSL23, WACCM, and MRIJMA). In broad terms, models with a stronger polar vortex than the reanalysis (see Fig. 1) have a lack of SSWs (GISSL23 and WACCM) and those with a weaker polar vortex than the reanalysis have an excess of SSWs (GISSL53, although not significantly so). The exception to this is MRIJMA, but the lack of SSWs here may be related to the climatological SST forcing.

b. *Climatology of warmings*

Figure 3 shows bar plots of the frequency of occurrence of SSWs in each month from November through March. The reanalysis (plotted in clear bars in each panel) shows the climatology from the NCEP–NCAR dataset, which is peaked in midwinter with much fewer events occurring in November and March. None of the GCMs examined in the present study reproduce this climatology of SSWs. None of the GCMs has its largest frequency of SSWs in January. FVGCM and GISSL53 are closest to the reanalysis data, both showing a distribution with its peak shifted toward February. The other GCMs show either a seasonality of events strongly skewed toward March (WACCM and MRIJMA) or an approximately flat distribution without obvious structure (GISSL23 and MAECHAM). It could be hypothesized that the large proportion of SSWs in November in MAECHAM is related to

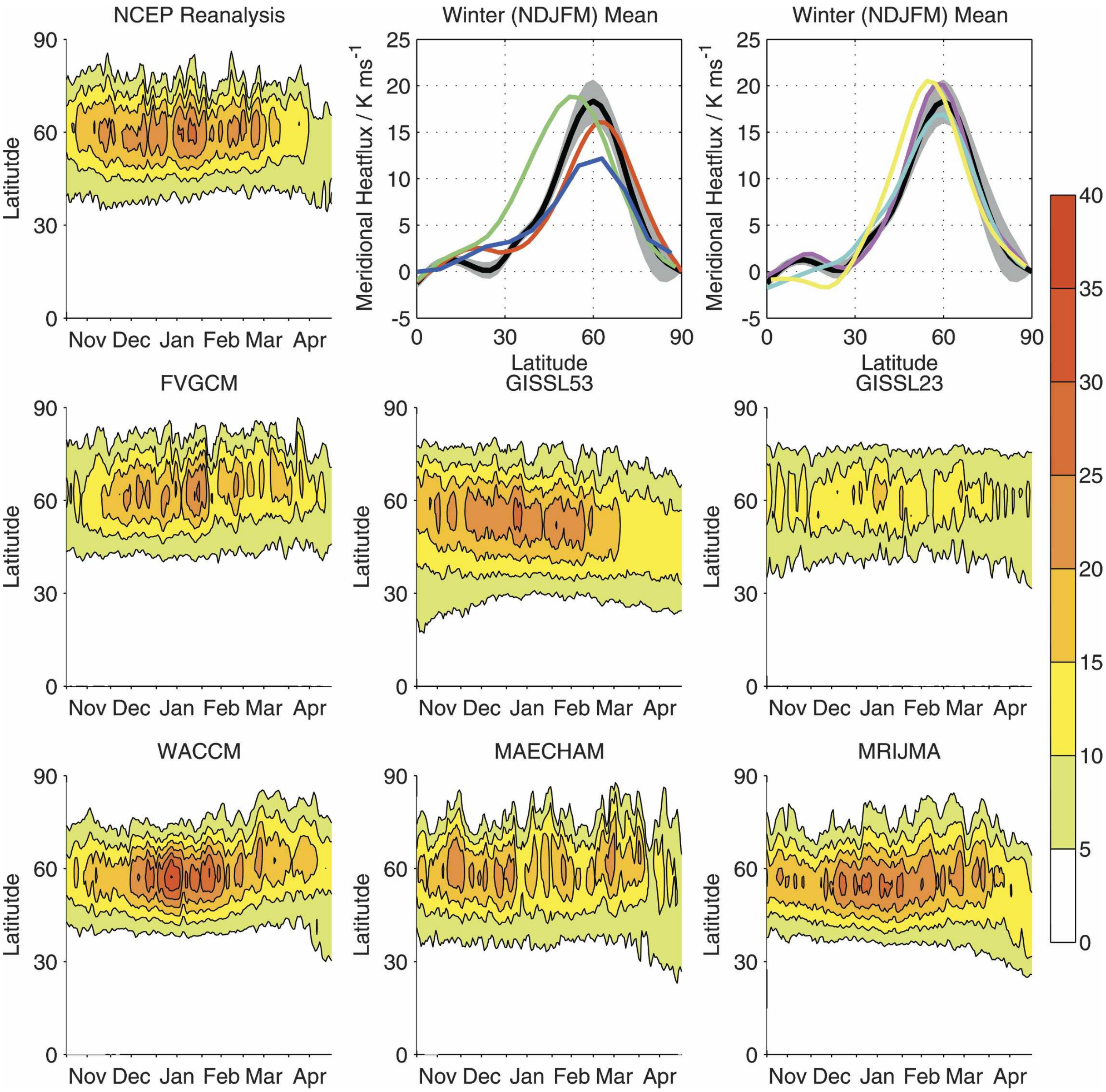

FIG. 2. Same as in Fig. 1, but for meridional heat flux climatology at 100 hPa. Contour interval is 5 K m s^{-1}.

MAECHAM's relatively weak zonal mean zonal winds during November (Fig. 1). Weak, climatological zonal mean zonal winds during February might also explain the large number of SSWs in GISSL53 during February. Two of the models with a lack of SSWs and a relatively strong zonal mean zonal wind jet at 10 hPa also have SSWs mostly toward the end of winter (WACCM and GISSL23).

c. *Type of warmings*

The type of SSWs in the GCM runs is shown in Table 4. In comparing the type of SSWs between the datasets we disregard the number of events and focus on the ratio between the number of vortex splits and vortex displacements. Column 2 shows the number of SSWs in each model, column 3 shows the number of vortex displacements, and column 4 shows the vortex splits. Column 5 shows the ratio between the number of vortex displacements and vortex splits. Columns 6 and 7 show if the distribution of vortex displacements and splits in each model is significantly different from the distribution in the NCEP–NCAR reanalysis. To statistically compare the type of events in each GCM, we construct bivariate tables of each GCM and the NCEP–NCAR

Table 3. Summary of SSWs in each GCM run.

GCM	Run length/winters	SSWs	Frequency/ events yr^{-1}	Standard error	Significant diff NCEP–NCAR at 0.10	Significant diff NCEP–NCAR at 0.05
FVGCM	49	23	0.47	0.09	No	No
GISSL53	47	37	0.79	0.08	No	No
GISSL23	46	12	0.26	0.07	Yes	Yes
WACCM	50	5	0.10	0.04	Yes	Yes
MAECHAM	29	15	0.52	0.10	No	No
MRIJMA	60	14	0.23	0.06	Yes	Yes
NCEP–NCAR	45	27	0.60	0.10		

reanalysis dataset and use an χ^2 test. For more details see appendix B.

There is wide variation between the GCMs in the type of SSW they produce. While FVGCM and WACCM show a ratio of events close to the reanalysis, the two GISS GCMs are biased toward splitting events, and MAECHAM and MRIJMA are biased toward displacement events. All of the GCMs that have a smaller proportion of vortex splitting events than the reanalysis datasets (FVGCM, WACCM, MAECHAM, and MRIJMA) are not significantly different from the NCEP–NCAR reanalysis at either the 0.10 or 0.05 confidence level. The two GISS GCMs have a larger proportion of vortex splitting events than either of the reanalysis datasets. GISSL53, which has the largest number of events, has significantly more vortex splitting events than the reanalysis at both the 0.10 and 0.05 confidence levels, while GISSL23 has significantly more events only at the 0.10 confidence level.

Part of the reason for this difference in ratio of SSW type could be the ratio of wavenumber 2 to wavenumber 1 planetary wave energy entering the stratosphere

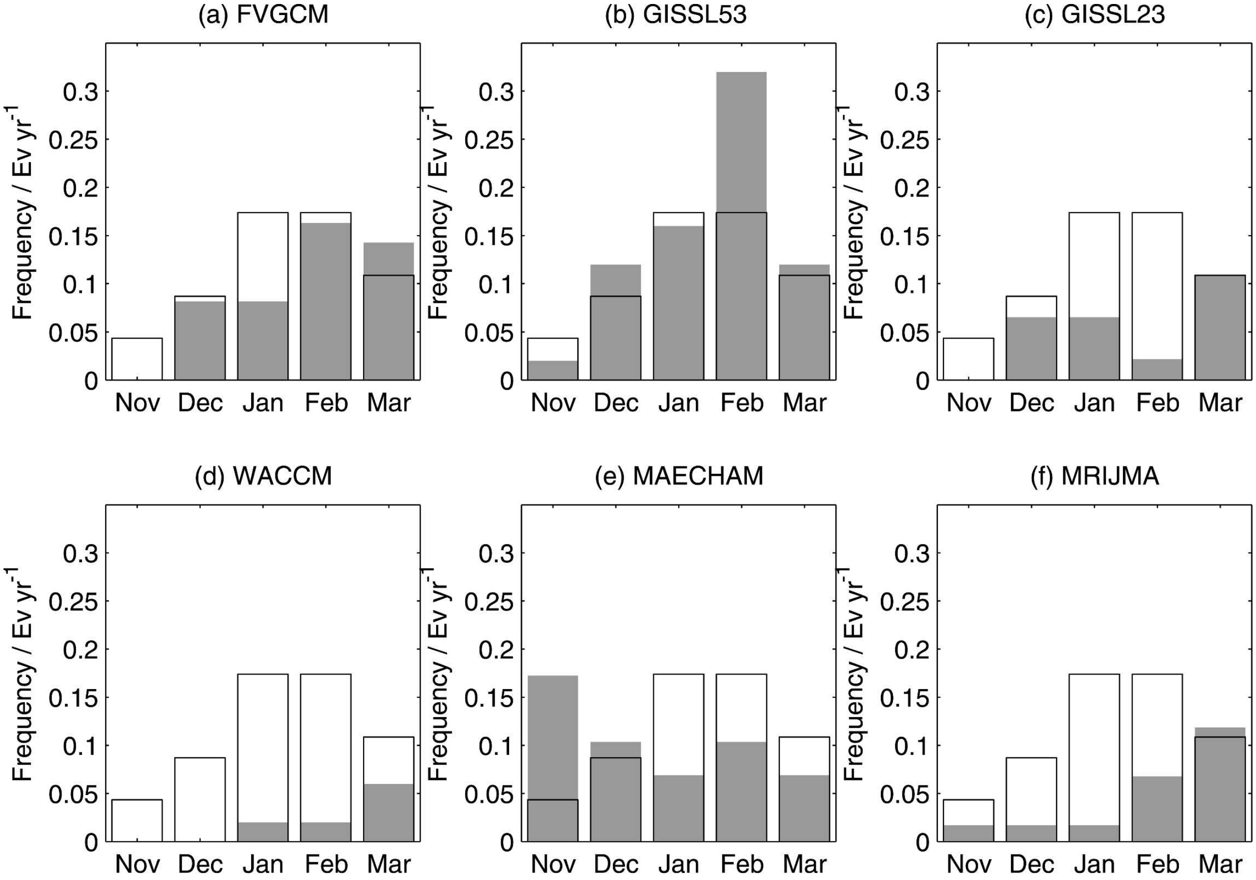

Fig. 3. SSW climatology in frequency of events per year in given month for each GCM, in gray bars: (a) FVGCM, (b) GISSL53, (c) GISSL23, (d) WACCM, (e) MAECHAM, and (f) MRIJMA. NCEP–NCAR reanalysis climatology is shown in unfilled bars.

TABLE 4. Type of SSWs in each GCM run.

GCM	SSWs	Vortex displacements	Vortex splits	Ratio	Significant diff NCEP–NCAR at 0.10	Significant diff NCEP–NCAR at 0.05	Ratio climatological heat flux $m = 1/m = 2$
FVGCM	23	13	10	1.3	No	No	0.96
GISSL53	37	11	26	0.4	Yes	Yes	0.61
GISSL23	12	3	9	0.3	Yes	No	1.18
WACCM	5	3	2	1.5	No	No	1.00
MAECHAM	15	10	5	2.0	No	No	0.57
MRIJMA	14	11	3	3.7	No	No	0.37
NCEP–NCAR	27	15	12	1.3			0.66

in each GCM. Column 8 in Table 4 shows the climatological ratio of area-weighted, winter-mean meridional heat flux between 45° and 75°N at 100 hPa due to wavenumber 2 and due to wavenumber 1 in each of the GCMs and the NCEP–NCAR reanalysis dataset. MRIJMA, which has a large bias toward vortex displacement events, also has a large bias toward wavenumber 1 heat flux compared to the reanalysis, while GISSL23, which has a large bias toward vortex splitting events also has a large bias toward wavenumber 2 heat flux compared to the reanalysis. However, for the other GCMs this pattern does not hold and the reasons for the SSW-type biases in the GCMs seem more complicated than the initial hypothesis expressed here.

5. Process-based performance of GCMs

In this section the dynamical benchmarks for SSWs established in CP06 are calculated for the SSWs found in each of the GCM runs in the present study. In all figures in this section, the distribution of benchmarks is shown with a box plot. The box of each plot shows the interquartile range. The central line of the box shows the median. The whiskers of the box show the minimum and maximum points in the distribution that are not outliers. Outliers are marked with an "x" and are defined as any points that are greater than 3/2 times the interquartile range from the ends of the box. The mean value of the diagnostic is shown by a cross. If the mean of the diagnostic in the GCM is significantly different at 95% confidence from the mean of the diagnostic in the reanalysis dataset, the mean is plotted with a filled circle. The mean of the diagnostic in each GCM and the reanalysis is compared with a standard two-sample t test with unequal variances [see appendix A, Eqs. (A3)–(A4)].

a. Amplitude in middle stratosphere (ΔT_{10})

The first benchmark is the area-weighted mean 10-hPa polar cap temperature anomaly, at 90°–50°N, ±5 days from the onset date (ΔT_{10}). The anomaly ΔT_{10} gives an indication of the amplitude of SSWs in each GCM. Figure 4 shows the distribution of ΔT_{10} for the NCEP–NCAR reanalysis and for the GCMs in the present study. The mean value of ΔT_{10} in the NCEP–NCAR reanalysis is 7.4 K. In all GCMs, the mean value of ΔT_{10} is not significantly different from the NCEP–NCAR reanalysis. The distribution of ΔT_{10} in the GCMs is symmetric and has a similar spread to the NCEP–NCAR reanalysis in all cases except WACCM. It is unclear, however, if this is a feature of the dynamics of WACCM or a consequence of the small number of SSWs in that model.

b. Amplitude in lower stratosphere (ΔT_{100})

The second benchmark is the area-weighted mean 100-hPa polar cap temperature anomaly, 90°–50°N, ±5 days from the onset date (ΔT_{100}). The quantity ΔT_{100} gives an indication of the strength of downward propa-

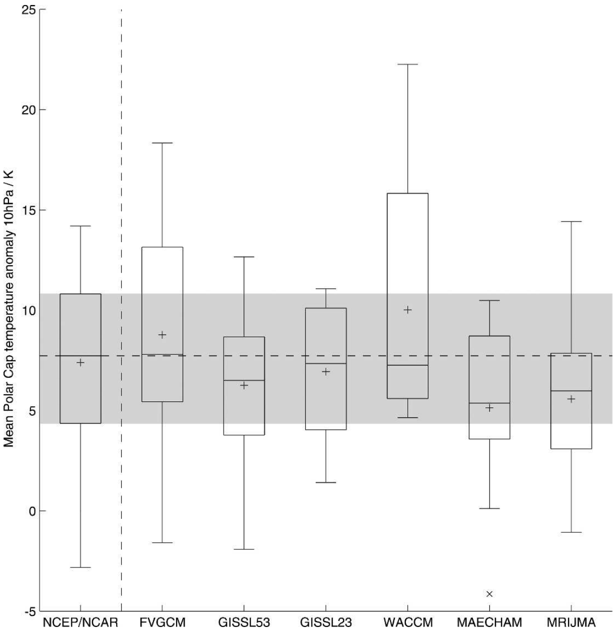

FIG. 4. Box plots showing distribution of maximum polar cap temperatures at 10 hPa (90°–50°N) for SSW in a range of GCMs and NCEP–NCAR reanalysis. Gray shading shows interquartile range of reanalysis; dashed black line shows median of reanalysis. Outliers are marked by "x." Mean is shown by a cross.

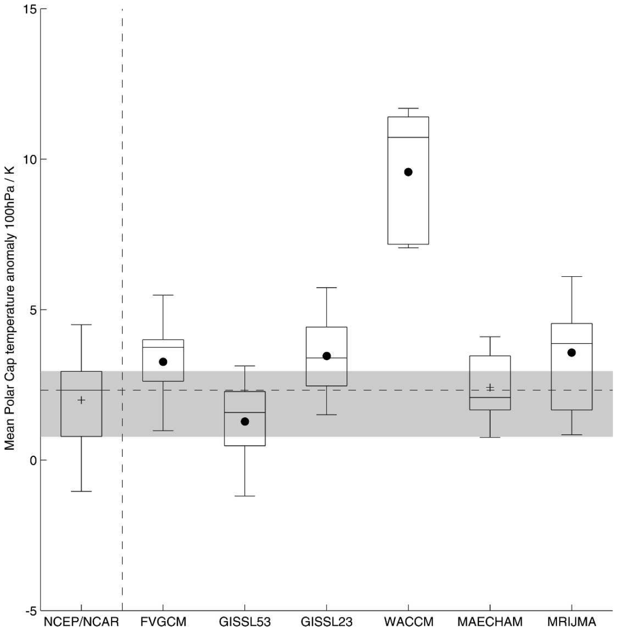

FIG. 5. Same as in Fig. 4, but at 100 hPa. The mean of each dataset is shown by a cross; if the mean of a GCM dataset is significantly different from the mean of the reanalysis, the cross is replaced by a filled circle.

gation of temperature anomalies in the stratosphere of each GCM. Figure 5 shows the distribution of ΔT_{100} for the NCEP–NCAR reanalysis and for the GCMs in the present study. The mean value of ΔT_{100} in the NCEP–NCAR reanalysis is 2.0 K. In all but one of the GCMs (MAECHAM) ΔT_{100} is significantly different from the polar cap temperature anomaly in the NCEP–NCAR reanalysis. GISSL23, FVGCM, WACCM, and MRIJMA all have significantly larger ΔT_{100} associated with SSWs. As discussed in more detail below, WACCM has particularly large temperature anomalies on this pressure surface. GISSL53 has significantly smaller ΔT_{100} than the NCEP–NCAR reanalysis.

c. *Deceleration of polar vortex jet* (ΔU_{10})

The third benchmark is the difference in zonal mean zonal wind, at 60°N and 10 hPa, 15–5 days prior to the onset date minus 0–5 days after the onset date (ΔU_{10}). The quantity ΔU_{10} gives an indication of the momentum deposition that accompanies SSWs. Figure 6 shows the distribution of ΔU_{10} for the NCEP–NCAR reanalysis and for the GCMs in the present study. The mean value of ΔU_{10} in the NCEP–NCAR reanalysis is 26.2 m s^{-1}. In all but one of the GCMs (GISSL53) ΔU_{10} is not significantly different from ΔU_{10} in the NCEP–NCAR reanalysis. GISSL53 has a significantly weaker ΔU_{10} than the NCEP–NCAR reanalysis. All of the GCMs, even those with numbers of SSWs comparable to the reanalysis, have a significantly greater interquartile range in ΔU_{10} than the NCEP–NCAR reanalysis.

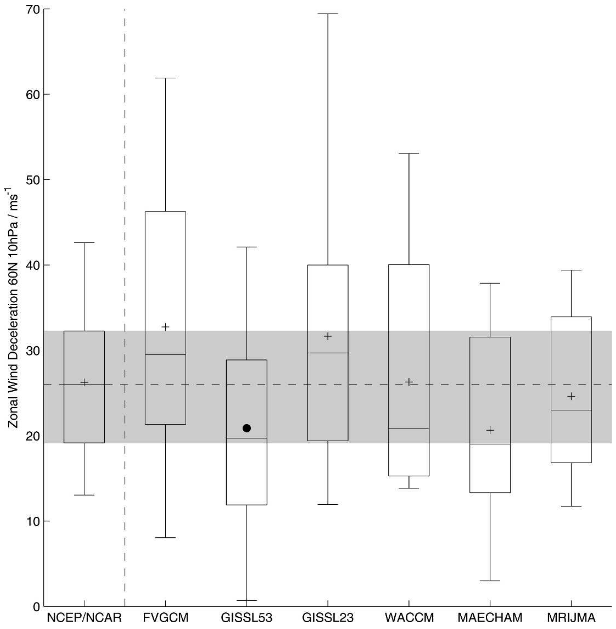

FIG. 6. Same as in Fig. 5, but for mean zonal wind deceleration at 60°N and 10 hPa.

d. *Heat flux anomaly* ($\Delta\overline{v'T'_{100}}$)

The fourth benchmark is the area-weighted, mean 100-hPa $\overline{v'T'}$ anomaly, 20–0 days before the onset date ($\Delta\overline{v'T'_{100}}$). The variable $\Delta\overline{v'T'_{100}}$ gives an indication of the input of planetary wave activity required to cause the SSWs in each GCM. Figure 7 shows the distribution of $\Delta\overline{v'T'_{100}}$ for the NCEP–NCAR reanalysis and for the GCMs in the present study. The mean value of $\Delta\overline{v'T'_{100}}$ in the NCEP–NCAR reanalysis is 8.5 K m s^{-1}. In all but one of the GCMs (GISSL53) $\Delta\overline{v'T'_{100}}$ is not significantly different from $\Delta\overline{v'T'_{100}}$ in the NCEP–NCAR reanalysis. GISSL53 has a significantly weaker $\Delta\overline{v'T'_{100}}$ than the NCEP–NCAR reanalysis. The median value of $\Delta\overline{v'T'_{100}}$ for all of the GCMs is smaller than the median value in the NCEP–NCAR reanalysis.

e. *A note on the duration of events in GISSL53*

In three of the four benchmarks presented in this section, GISSL53 produces SSWs that are significantly different from those found in the NCEP–NCAR reanalysis. To investigate the reasons for this difference it proves useful to examine the duration of SSWs. To compute the duration of events in NCEP–NCAR reanalysis and each GCM we count the number of consecutive days of easterly zonal mean zonal winds at 60°N and 10 hPa after the onset date of each SSW in each model. SSWs are grouped into three categories.

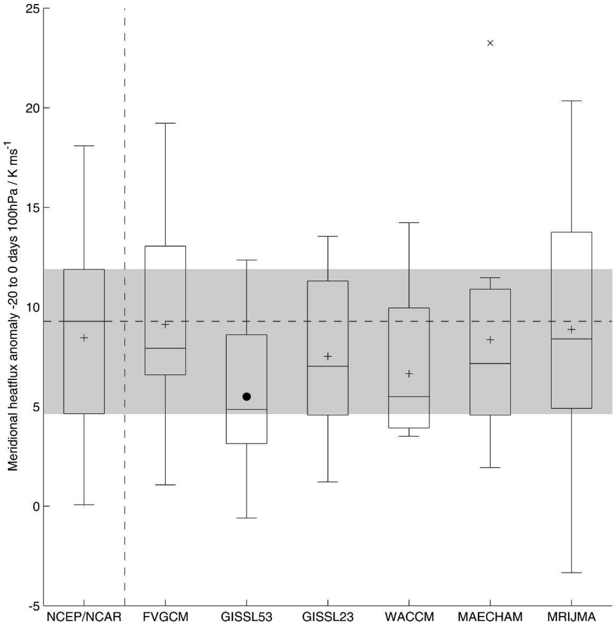

FIG. 7. Same as in Fig. 5, but for mean meridional heat flux anomaly (75°–45°N) −20 to 0 days prior to SSWs.

Table 5 shows the number of SSWs in each GCM with duration less than 4 days (column 2) and 4–9 days (column 3), and with duration longer than 9 days (column 4). The distribution of events in each category can be compared with the NCEP–NCAR reanalysis using an χ^2 test (appendix B); the results of this test at the 0.05 confidence level are shown in column 5. The mean duration of SSWs in each GCM is shown in column 6.

Table 5 shows that GISSL53 and MAECHAM have a significantly different distribution of SSW duration than the NCEP–NCAR reanalysis. GISSL53 has a higher proportion of short duration events than the NCEP–NCAR reanalysis (and also a distinct lack of events of the 4–9-day length). The majority of these short duration SSWs occur during February (9 of 20) in GISSL53, whereas they are fairly evenly distributed in the NCEP–NCAR reanalysis. The GISSL53 model has an unusually weak mean stratospheric vortex strength during February, which might be related to frequent SSWs in this month (Fig. 1).

To test if the high proportion of short duration (less than 4 days) SSWs in GISSL53 biases the distribution of benchmarks and distorts the results in the previous section, all of the benchmarks are recomputed, excluding events in the short event category. Figure 8 compares each of the dynamical benchmarks in the NCEP–NCAR reanalysis and GISSL53 models with and without short-duration SSWs. For both ΔU_{10} and $\Delta\overline{v'T'_{100}}$, the difference between the mean benchmark for the NCEP–NCAR reanalysis and GISSL53 is reduced when short events are excluded. We conclude that the short duration events, which are unusually prevalent in GISSL53, bias the benchmark diagnostics in the previous section.

6. Discussion

From the previous two sections we have arrived at two different but not contradictory conclusions about SSWs in the GCMs described here. First, in section 4 we showed that three of the six GCM simulations of the winter stratosphere produce a smaller number of SSWs than are observed in reanalysis datasets. Second, we showed in section 5 that the GCMs in the present study produce SSWs whose characteristics are statistically similar to SSWs observed in the NCEP–NCAR reanalysis. This suggests that the lack of SSWs observed in section 4 is not due to an inability of the GCMs to reproduce the dynamics of SSW events but that the dynamics are produced less frequently than in the reanalysis dataset.

With this in mind, in this section we attempt to reconcile these two results by asking a number of questions of the GCM simulations, designed to try and understand some of the reasons for the low frequency of SSWs simulated by GISSL23, WACCM, and MRIJMA.

a. Is the relationship between $\Delta\overline{v'T'_{100}}$ and ΔT_{10} *correct in the GCMs?*

Previous studies have shown that there is a good relationship between the wintertime average meridional

TABLE 5. Duration of SSWs in each GCM.

Model	Short events <4 days	Normal events 4–9 days	Long events >9 days	Significant diff NCEP–NCAR at 0.05	Mean duration/days
FVGCM	6 (26.1%)	10 (43.5%)	7 (30.4%)	No	7.2
GISSL53	20 (54.1%)	7 (18.9%)	10 (27.0%)	Yes	6.1
GISSL23	5 (41.7%)	5 (41.7%)	2 (16.7%)	No	5.4
WACCM	2 (40.0%)	1 (20.0%)	2 (40.0%)	No	8.2
MAECHAM	3 (20.0%)	2 (13.3%)	10 (66.7%)	Yes	11.8
MRIJMA	4 (28.6%)	6 (42.9%)	4 (28.6%)	No	8.0
NCEP–NCAR	12 (44.5%)	13 (48.1%)	2 (7.4%)		5.2

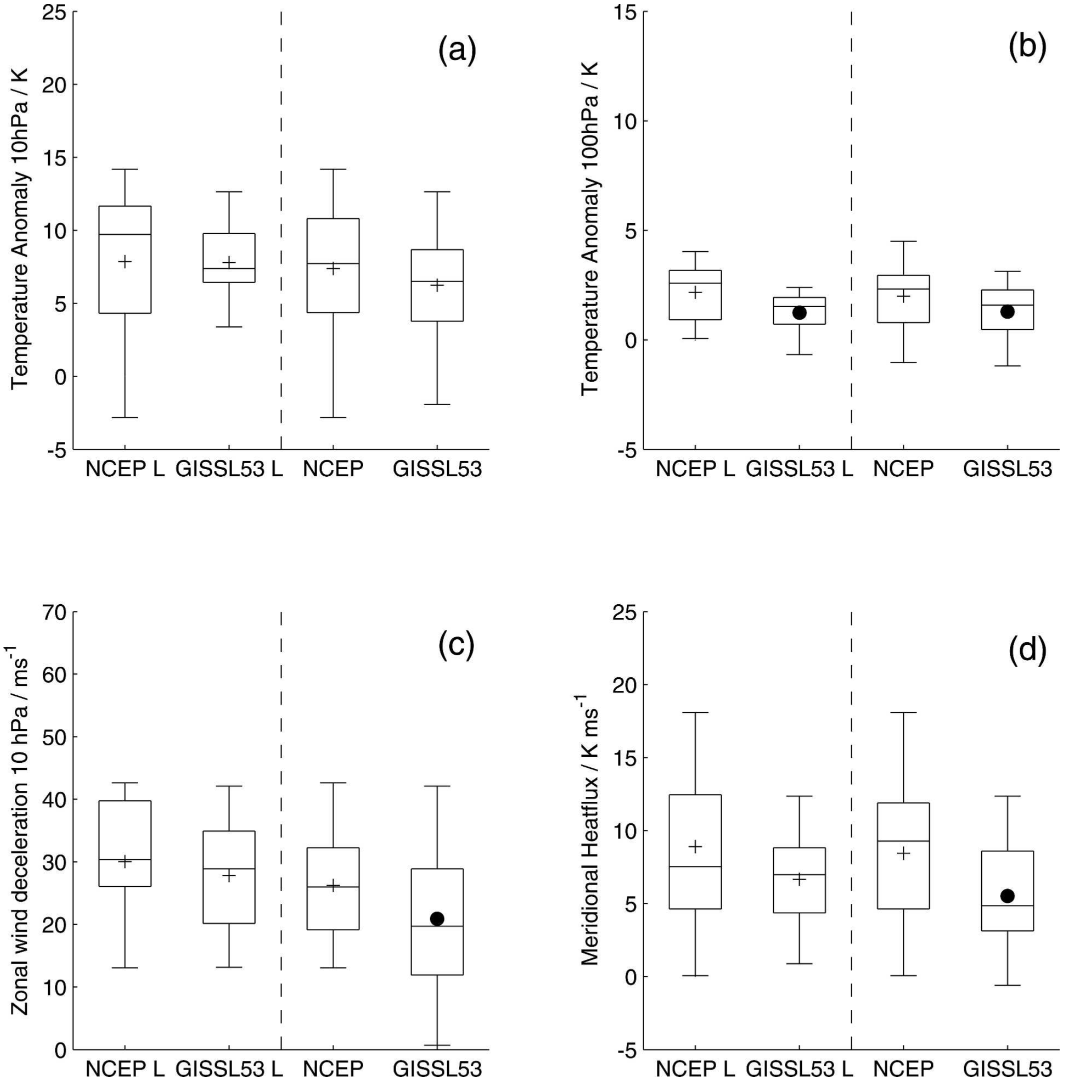

FIG. 8. Box plots of each of the four dynamical benchmarks for SSWs in NCEP–NCAR reanalysis and GISSL53 excluding short-duration events (NCEP–NCAR L and GISSL53 L) and for all events (NCEP–NCAR and GISSL53). (a) Temperature anomaly at 10 hPa, (b) temperature anomaly at 100 hPa, (c) zonal mean zonal wind deceleration, and (d) meridional heat flux anomaly. For details of box plots and benchmarks see previous sections.

heat flux at 100 hPa and the wintertime average polar cap temperatures in the middle stratosphere (Newman et al. 2001; Hu and Tung 2002). In this section we examine if a similar relationship exists for the SSWs in the NCEP–NCAR reanalysis and the GCM runs in this study.

Figure 9 shows $\Delta\overline{v'T'_{100}}$ plotted against ΔT_{10} (the first and fourth benchmarks discussed in the previous section). SSWs in the NCEP–NCAR reanalysis are plotted in black dots, and the corresponding linear regression of the two quantities is plotted in the black line. There is an obvious linear relationship between the two quantities. Each of the panels in Fig. 9 shows the same quantities plotted for SSWs in each of the GCM runs. Most of the GCMs show a relationship close to that observed in the NCEP–NCAR reanalysis. Only GISSL53 and MAECHAM show a large spread in the relationship between meridional heat flux anomaly and polar cap temperature anomaly.

To assess if the relationship between $\Delta\overline{v'T'_{100}}$ and ΔT_{10} for SSWs in the NCEP–NCAR reanalysis is replicated by the GCMs we compare the value of the regression coefficient in each GCM with the reanalysis using a *t* test. Details of the regression equations and the estimation of their standard error are shown in appendix C. Table 6 shows the correlation coefficient between $\Delta\overline{v'T'_{100}}$ and ΔT_{10} (column 1), the corresponding regression coefficient between $\Delta\overline{v'T'_{100}}$ and ΔT_{10} (col-

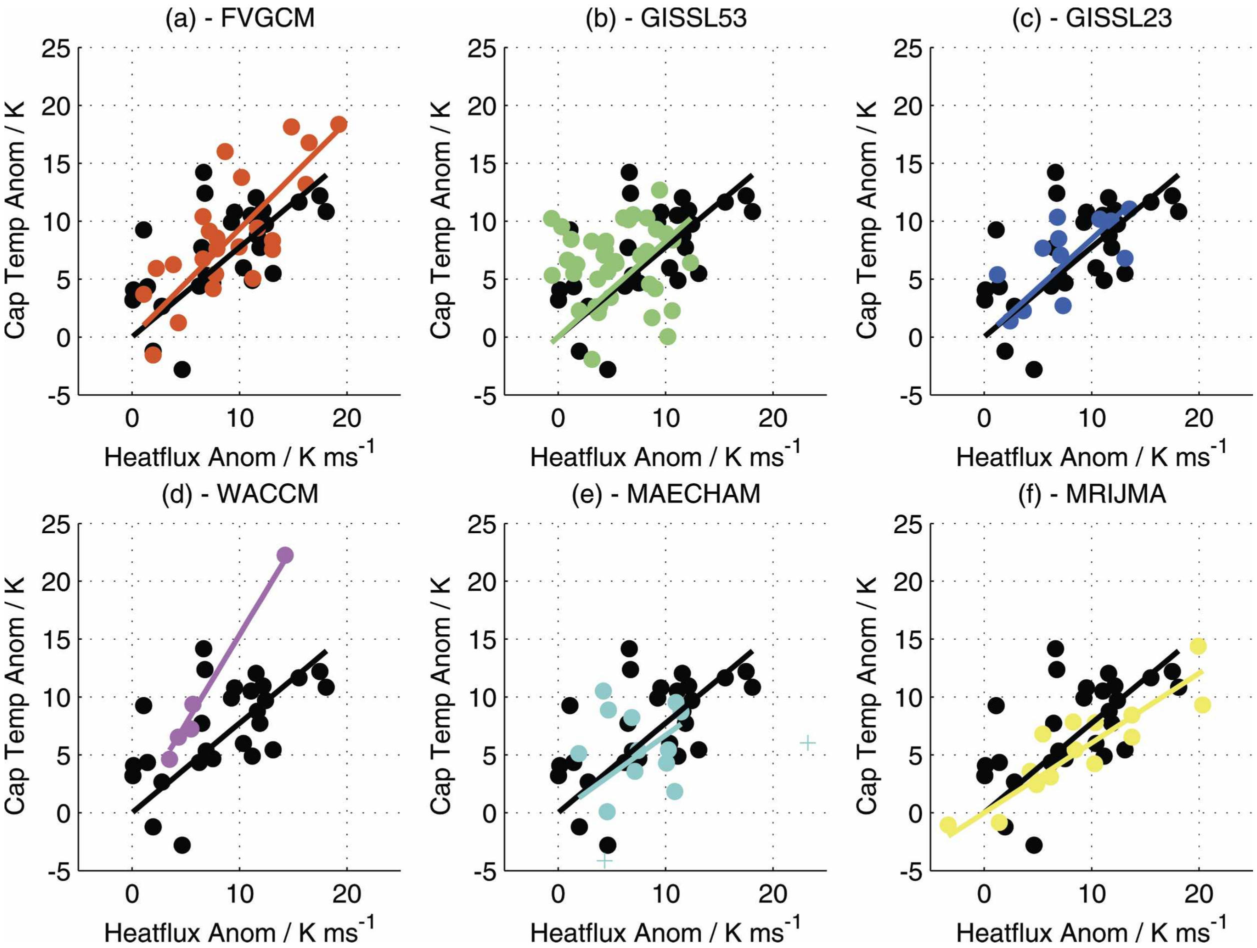

FIG. 9. Scatterplot showing maximum polar cap temperature anomaly vs integrated meridional heat flux anomaly for each SSW in NCEP–NCAR reanalysis and analyzed GCM runs. Colored lines show linear regression for each dataset. Crosses in MAECHAM panel indicate outlier points that are excluded from the regression and correlation calculations.

umn 2) and the coefficient of determination (column 3) and the standard error (column 4) for the regression. Columns 5 and 6 show if the regression coefficient in each GCM is significantly different from the regression coefficient in the NCEP–NCAR reanalysis at the 0.10 and 0.05 confidence levels.

As was indicated by visual inspection, four of the GCMs (FVGCM, GISSL23, WACCM, and MRIJMA) have a high degree of correlation between $\Delta\overline{v'T'_{100}}$ and ΔT_{10} and hence a good fit indicated by the coefficient of determination. Of the GCMs here, only one (WACCM) has a significantly different relationship between $\Delta\overline{v'T'_{100}}$ and ΔT_{10}. Both GISSL23 and MRIJMA have statistically similar relationships between $\Delta\overline{v'T'_{100}}$

TABLE 6. Relationship between heat flux and polar cap temperature anomaly for SSWs in each GCM.

GCM	Correlation	Regression coef	R^2	Significant diff NCEP–NCAR at 0.10	Significant diff NCEP–NCAR at 0.05
FVGCM	0.76	0.93	0.75	No	No
GISSL53	0.04	0.82	0.77	No	No
GISSL23	0.66	0.84	1.04	No	No
WACCM	0.99	1.54	0.89	Yes	Yes
MAECHAM	0.02	0.67	0.42	No	No
MRIJMA	0.89	0.60	0.94	No	No
NCEP–NCAR	0.59	0.77	0.86		

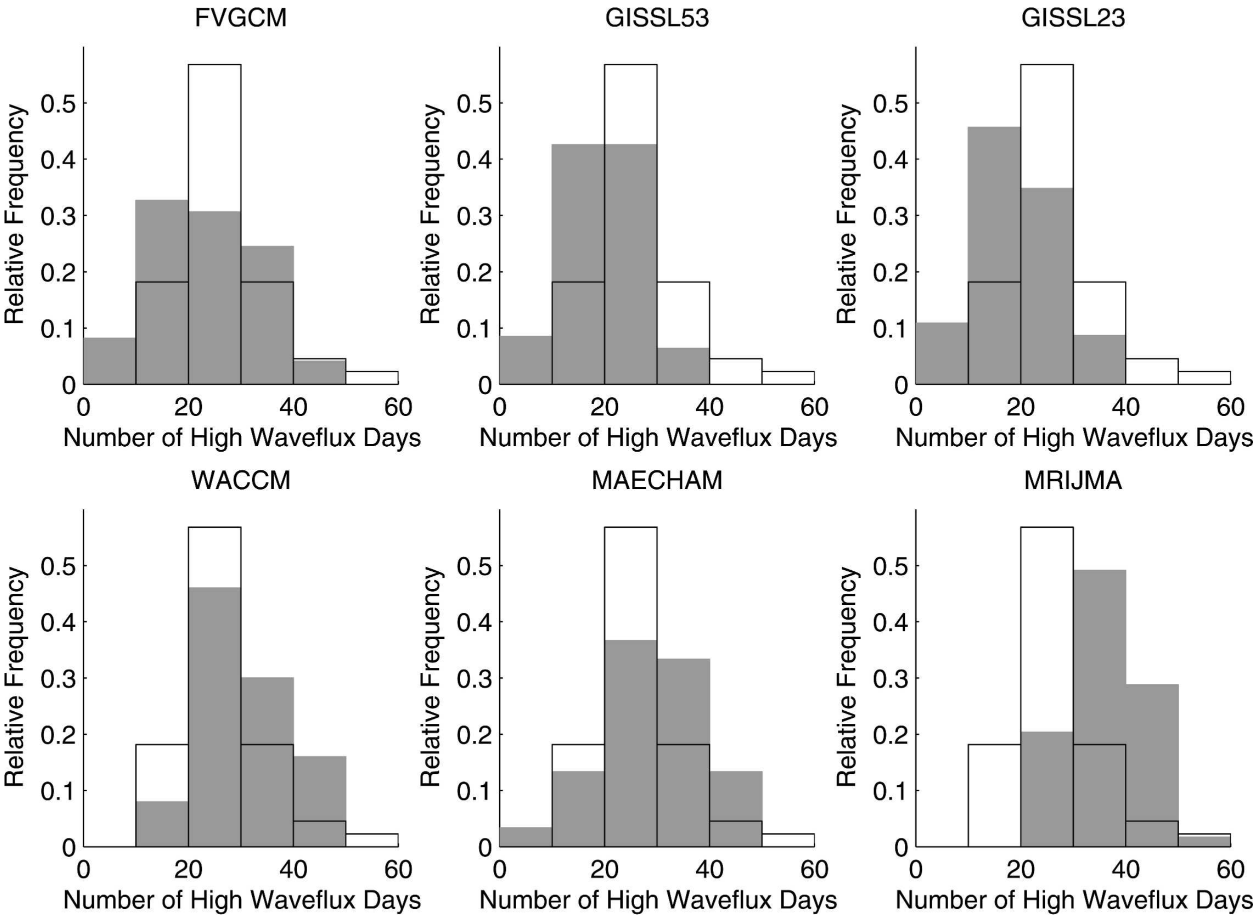

FIG. 10. Histogram showing number of days of heat flux anomaly greater than 8.5 K m s^{-1} in NDJFM each winter. Gray bars show each GCM; white bars show NCEP–NCAR reanalysis.

and ΔT_{10} as the NCEP–NCAR reanalysis, suggesting that a weak response of the stratosphere to meridional heat flux anomalies is not the cause of the significant lack of warmings in either model.

There are however some caveats to these results; as noted before, the sample size of the WACCM diagnostic is relatively small. It is also difficult to have much confidence in the MAECHAM and GISSL53 results, given that the correlation between $\Delta\overline{v'T'_{100}}$ and ΔT_{10} in these GCMs is very small. The correlation between $\Delta\overline{v'T'_{100}}$ and ΔT_{10} increases to 0.29 for GISSL53 when events of short duration (less than 4 days) are excluded and to 0.20 for MAECHAM when events of long duration (more than 9 days) are excluded. These correlations are still much less than the equivalent for the NCEP–NCAR reanalysis.

b. Do the GCMs simulate the correct number of days of extreme $\Delta\overline{v'T'_{100}}$?

Another reason for the lack of SSW activity in some of the GCMs, might be that the frequency of extreme meridional heat flux anomalies, which tend to precede SSWs as seen in the previous section, is lower than that of the reanalysis data. To assess this we count the number of days with extreme meridional heat flux anomalies in each GCM. For each winter season [November–March (NDJFM)] the number of days with a mean area-weighted meridional heat flux anomaly greater than 8.5 K m s^{-1} [the mean value of $\Delta\overline{v'T'_{100}}$ for SSWs in the NCEP–NCAR reanalysis (cf. Fig. 7)] is calculated.

Figure 10 shows histograms for each of the GCMs in gray bars and for the NCEP–NCAR reanalysis in the clear bars. Only the two GISS GCMs have distributions with a noticeable lack of extreme heat flux anomalies compared to the reanalysis. This is particularly true of GISSL23; GISSL23 also has a mean heat flux much smaller than the reanalysis data, indicating that it has a distinct lack of large absolute heat flux as expected. Two of the GCMs that have a significant lack of SSW activity (WACCM and MRIJMA) have distributions indicating a higher frequency of extreme heat flux days than the reanalysis.

TABLE 7. Number of extreme heat flux days per winter in each GCM.

GCM	Expected extreme heat flux days	Standard error	Significant diff NCEP–NCAR at 0.10	Significant diff NCEP–NCAR at 0.05
FVGCM	24.4	1.3	No	No
GISSL53	21.0	0.9	Yes	Yes
GISSL23	20.3	1.0	Yes	Yes
WACCM	30.9	1.1	Yes	Yes
MAECHAM	30.1	1.8	No	No
MRIJMA	35.9	0.9	Yes	Yes
NCEP–NCAR	26.8	1.3		

The expected number of extreme heat flux days during each winter season in the reanalysis and in each GCM can be calculated and compared using a *t* test (see appendix A). The results of this analysis are shown in Table 7. All of the GCMs apart from FVGCM and MAECHAM have a significantly different number of extreme heat flux days compared to the NCEP–NCAR reanalysis. Particularly extreme are GISSL23, which has only approximately 75% the number of extreme heat flux days as the NCEP–NCAR reanalysis, and MRIJMA, which has approximately 134% the number of extreme heat flux days as the NCEP–NCAR reanalysis.

7. Conclusions

In this paper, we have attempted to use a new climatology of SSWs, developed by CP06, to evaluate the performance of several well-documented and widely used general circulation models of the middle atmosphere. The overall results of the study are encouraging: a number of the GCMs compare well with the climatology. A summary of the results of the study is shown in Table 8. Columns two and three ask the questions: Does the GCM simulate the correct number of SSWs? And does the GCM simulate the correct ratio of vortex splitting and vortex displacement SSWs? The middle section of the table (columns 4–8) shows the process-based benchmarks for SSWs introduced in CP06. Columns are left blank if there is no significant difference between the diagnostic in the particular GCM and the NCEP–NCAR reanalysis. The final section of the table shows the two further diagnostics of the GCMs dynamical behavior discussed in section 6.

Of the GCMs studied, FVGCM and MAECHAM performed well in both the climatological and process-based comparisons of SSWs with the NCEP–NCAR reanalysis data. GISSL53 performed well in the climatological comparison of SSWs, and in the process-based comparisons when short duration events were excluded. GISSL53 and MAECHAM, however, fail to show a high correlation between the anomalous heat flux prior to the SSW and the polar cap temperature in the middle stratosphere during the warming.

The other three GCMs in the study (GISSL23, WACCM, and MRIJMA) have stratospheres that, at least in terms of major midwinter warming activity, are too quiescent. Nevertheless, the small numbers of major warming events produced by these GCMs compare well in a number of process-based benchmarks with the SSWs observed in the NCEP–NCAR reanalysis.

The diagnostics presented in section 6 give some clues as to the reason for the lack of warming activity in GISSL23, WACCM, and MRIJMA. GISSL23 has a marked lack of meridional heat flux both in the mean and variability, which suggests that a lack of disturbance by tropospheric Rossby waves is the reason for its lack of SSW activity. MRIJMA has neither a significantly weaker relationship between the heat flux and polar cap temperature anomaly nor significantly fewer extreme heat flux anomaly days than the NCEP–NCAR reanalysis. One obvious deficiency of the MRIJMA is its lack of a zonal wavenumber 2 heat flux (see Table 3), and its apparent lack of vortex splitting SSWs. Further statistical tests would be required to confirm this hypothesis. We suggest that the lack of a wavenumber 2 heat flux might be related to the climatological SST conditions used in the MRIJMA run, which would not include strong ENSO events.

WACCM is the most surprising of the six models studied here; it does not appear to be deficient in the mean and variability of meridional heat flux in the lower stratosphere or in the relationship between meridional heat flux and polar cap temperature during SSWs. The only significant difference between the GCM and the reanalysis is in its simulation of the zonal mean zonal jet, which is much too strong, and in the temperature anomaly in the lower stratosphere during SSWs, which is also too large. We might speculate that the very strong vortex in WACCM means that tropospheric Rossby wave activity cannot propagate into the middle stratosphere easily, and that evidence of this is

Table 8. Summary of characteristics of each GCM.

GCM	Frequency	Type	ΔT_{10}	ΔT_{100}	ΔU_{10}	$\Delta\overline{v'T'_{100}}$	Heat flux vs cap temperature	Extreme heat flux
FVGCM	Yes	Yes		Large				
GISSL53	Yes	No		Small				Small
GISSL23	No	No		Large				Small
WACCM	No	Yes		Large			Large	Large
MAECHAM	Yes	Yes						
MRIJMA	No	Yes		Large				Large

provided by the high lower-stratospheric temperature during the major warming. In other words, much of the momentum deposition from planetary wave activity in WACCM may occur below the 10-hPa pressure surface where we have defined SSWs. It should be kept in mind however that the interpretation of the WACCM results requires some caution because of the small number of SSWs found in the simulation studied.

Finally, there appears to be some relationship between the strength of the climatological vortex in the middle stratosphere and the number of SSWs. In CP06 we showed that SSWs make little impact on wintertime mean polar cap temperatures (CP06, their Fig. 5). It is therefore appropriate to conclude that the climatology is largely unaffected by the frequency of SSWs. On the other hand, in the present study we have found that those GCMs with a strong climatological vortex also tend to have a lower frequency of SSWs than the reanalysis, and those GCMs with a weak climatological vortex also tend to have higher-frequency SSWs than the reanalysis. Thus, the strength of the climatological vortex might be a useful guide to the ability of GCMs to produce the observed frequency of SSWs.

One of the key difficulties in reproducing the observed climatological vortex is the tuning of the gravity wave parameterization (GWP) in stratosphere-resolving GCMs, and in particular the gravity wave momentum deposition in the mesosphere (Manzini and McFarlane 1998). The effect of tuning the GWP on SSW frequency has been little studied, although there is some evidence that the variability of the stratosphere may be insensitive to changes to the GWP (Christiansen 1999), and this would be an interesting extension to this paper when and if appropriate runs with a single GCM and a variety of GWP setups becomes available. We hope that the climatological and process-based benchmarks for the simulation of SSWs introduced in this study will provide an additional constraint that will prove useful to modelers wishing to tune stratosphere-resolving GCMs. In this study we also restricted our analysis to major SSWs (as defined by CP06); it might be interesting and useful in the future to investigate the more frequent minor warming activity present in GCMs and reanalysis.

The study shows that making comparisons between GCMs and their simulation of major warming activity is both useful and illuminating. Analyzing the variability of the northern winter stratosphere is important to understanding northern polar ozone chemistry, future changes to northern winter climate, and the impact of these changes on the troposphere. This study shows that certain stratosphere-resolving GCMs might not be suitable for analyzing these three important issues, given that they fail to simulate observed stratospheric variability successfully. While no common solution to the deficiencies identified in the GCMs immediately arises, there is nonetheless a great deal of progress that can be made by considering very simple diagnostics.

Acknowledgments. AJC and LMP were funded by an award to the Cooperative Institute for Climate Applications and Research (CICAR) from the U.S. National Oceanic and Atmospheric Administration and by a grant to Columbia University from the U.S. National Science Foundation. JP's research is funded by the NASA Climate Program Office. NCEP–NCAR reanalysis data were obtained from the Ingrid data server at Columbia University. We thank Jeff Jonas and Patrick Lonergan for carrying out the GISSL53 and GISSL23 runs, respectively. The paper greatly benefited from discussion with Darryn Waugh, Gavin Esler, and Matthew Wittman.

APPENDIX A

Statistical Test of SSW Frequency

To compare the frequency of SSWs in each GCM run with the frequency of SSWs in the reanalysis data we consider each winter in the reanalysis or model run to be a separate, independent observation of the frequency of major warming events per winter. Thus, for example, in the NCEP–NCAR dataset we have 45 observations of the frequency of events per winter, 23 observations with no events, 17 with one event, and 5 with two events. The sample mean frequency ($\overline{x}$) of SSWs per winter season is the expected value of the 45 observations in the NCEP–NCAR reanalysis:

$$\overline{x} = E[X] = \sum_{x} x\text{Pr}\{X = x\}. \tag{A1}$$

In Eq. (A1) x represents an observed frequency of SSWs per winter and $\text{Pr}\{X = x\}$ is the probability with which that frequency is observed. The sample variance of the frequency of warming events s^2 is calculated in a similar fashion using

$$s^2 = \sum_{x} (x - \overline{x})^2\text{Pr}\{X = x\}. \tag{A2}$$

The sample variance and the number of SSWs in each dataset N are used to estimate the sample standard error e of the expected mean frequency of SSWs:

$$e = \frac{\sqrt{s^2}}{\sqrt{N}}. \tag{A3}$$

The values of $\overline{x}$ and e are used to construct a t test that compares the mean frequency of SSWs in the reanalysis and a given GCM. The null hypothesis of this test is as follows: *The mean frequency of SSWs in the GCM and NCEP–NCAR reanalysis is equal*. The test is two-sided because there is no a priori reason to expect that the difference between the means should be positive or negative. A t statistic comparing the expected frequency of each model run and the NCEP–NCAR reanalysis is calculated in the standard way (Wilks 1995, p. 122):

$$t = \frac{\overline{x}_r - \overline{x}_m}{[e_r^2 + e_m^2]^{1/2}}. \tag{A4}$$

The r subscript denotes reanalysis statistics and the m subscript denotes model statistics. The t statistic is then compared to critical values of a t distribution with degrees of freedom calculated using the expression,

$$df = \frac{(e_r^2 + e_m^2)^2}{(e_r^2)^2/(N_r - 1) + (e_m^2)^2/(N_m - 1)}. \tag{A5}$$

APPENDIX B

Statistical Test of SSW Type

To test if each GCM has a similar proportion of vortex displacements and splits we construct a nonparametric χ^2 test. The null hypothesis of this test is as follows: *The frequency distributions of vortex displacement and splitting events in the GCM and NCEP–NCAR reanalysis are the same*. First, a table is constructed showing the number of vortex displacements and splits in the GCM and the NCEP–NCAR reanalysis. For example, for the FVGCM model the table reads,

	Vortex displacements	Vortex splits
FVGCM	13	10
NCEP–NCAR	15	12

Second the expected number of observations, E_{ij} for each cell is calculated, where i is the row index and j is the column index. Under the null hypothesis that the samples are drawn from the same distribution,

$$E_{ij} = \left(\frac{\sum_{j=1}^{J} O_{ij}}{\sum_{i=1}^{I}\sum_{j=1}^{J} O_{ij}}\right)\sum_{i=1}^{I} O_{ij}, \tag{B1}$$

where O_{ij} is the observed number of SSWs in each box of the table, I is the total number of columns, and J is the total number of rows. The expected frequencies for Table B1 are

	Vortex displacements	Vortex splits
FVGCM	12.88	10.12
NCEP–NCAR	15.12	11.88

The χ^2 parameter is estimated using the equation,

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}. \tag{B2}$$

For the FVGCM example $\chi^2 = 0.0047$. The value of χ^2 can then be compared to critical values of the χ^2 distribution with one degree of freedom. Critical values at the 0.10 and 0.05 confidence levels are 2.7060 and 3.8410, respectively. In the FVGCM example, the χ^2 value is well below the critical values so we cannot reject the null hypothesis that the frequency distributions of vortex displacement and splitting events in FVGCM and the NCEP–NCAR reanalysis are the same.

APPENDIX C

Statistical Test of Regression Parameters

For SSWs in both the GCM runs and the NCEP–NCAR reanalysis the regression model is as follows:

$$y_i = \beta x_i + e_i, \tag{C1}$$

where y represents ΔT_{10}, x represents $\Delta\overline{v'T'_{100}}$, e_i represents normally distributed residuals, and β is the parameter of the model to be determined. No constant term is included in the regression equation, because it is assumed that when $\Delta\overline{v'T'_{100}}$ is equal to zero ΔT_{10} is also equal to zero. The value of β is estimated using the equation

$$\beta = \frac{\sum_{i=1}^{N} y_i x_i}{\sum_{i=1}^{N} x_i^2}. \tag{C2}$$

The sum of the squared residuals is calculated from the equation

$$r = \sum_{i=1}^{N} (y_i - \beta x_i)^2. \tag{C3}$$

The sum of the squared residuals is used to give an unbiased estimate of the residual variance, $s^2 = r/(n-1)$. The standard error of the slope parameter is defined by

$$e = \frac{s}{[\sum_{i=1}^{N} x_i^2]^{1/2}}. \tag{C4}$$

As in appendix A, a *t* test is constructed using Eqs. (A3) and (A4) to compare the slope parameter of each model and the NCEP–NCAR reanalysis. The null hypothesis of this test is as follows: *The regression parameter β is the same for SSWs in the model and the NCEP–NCAR reanalysis.*

REFERENCES

Andrews, D. G., J. R. Holton, and C. B. Leovy, 1985: *Middle Atmosphere Dynamics.* Academic Press, 489 pp.

Austin, J., and Coauthors, 2003: Uncertainties and assesments of chemistry-climate models of the stratosphere. *Atmos. Chem. Phys.,* **3,** 1–27.

Bloom, S., and Coauthors, 2005: Documentation and validation of the Goddard Earth Observing Ststem (GEOS) Data Assimilation System—Version 4. Technical Report Series on Global Modelling and Data Assimilation, Tech. Rep. 26, 187 pp.

Butchart, N., J. Austin, J. R. Knight, A. A. Scaife, and M. L. Gallani, 2000: The response of the stratospheric climate to projected changes in the concentrations of well-mixed greenhouse gases from 1992 to 2051. *J. Climate,* **13,** 2142–2159.

Charlton, A. J., and L. M. Polvani, 2007: A new look at stratospheric sudden warmings. Part I: Climatology and modeling benchmarks. *J. Climate,* **20,** 449–469.

Chiba, M., K. Yamazaki, K. Shibata, and Y. Kuroda, 1996: The description of the MRI atmospheric spectral GCM (MRI-GSPM) and its mean statistics based on a 10-year integration. *Papers Meteor. Geophys.,* **47,** 1–45.

Christiansen, B., 1999: Stratospheric vacillations in a general circulation model. *J. Atmos. Sci.,* **56,** 1858–1872.

Erlebach, P., U. Langematz, and S. Pawson, 1995: Simulations of stratospheric sudden warmings in the Berlin troposphere-stratosphere-mesosphere GCM. *Ann. Geophys.,* **14,** 443–463.

Eyring, V., and Coauthors, 2005: A strategy for process-oriented validation of coupled chemistry–climate models. *Bull. Amer. Meteor. Soc.,* **86,** 1117–1133.

Fortuin, J. P. F., and H. Kelder, 1998: An ozone climatology based on ozonesonde and satellite measurements. *J. Geophys. Res.,* **103,** 31 709–31 734.

Gates, W. L., and Coauthors, 1999: An overview of the results of the Atmospheric Model Intercomparison Project (AMIP I). *Bull. Amer. Meteor. Soc.,* **80,** 29–55.

Hansen, J., G. Russell, D. Rind, P. Stone, A. Lacis, S. Lebedeff, R. Ruedy, and L. Travis, 1983: Efficient three-dimensional global models for climate studies: Models I and II. *Mon. Wea. Rev.,* **111,** 609–662.

Hartmann, D. L., J. M. Wallace, V. Limpasuvan, D. W. J. Thompson, and J. R. Holton, 2000: Can ozone depletion and global warming interact to produce rapid climate change? *Proc. Natl. Acad. Sci. USA,* **97,** 1412–1417.

Hu, Y., and K. K. Tung, 2002: Interannual and decadal variations of planetary wave activity, stratospheric cooling, and Northern Hemisphere Annular Mode. *J. Climate,* **15,** 1659–1673.

JMA, 1997: Outline of the operational numerical weather prediction at the Japan Meteorological Agency. Japan Meteorological Agency, 120 pp.

Kallberg, P., A. Simmons, S. Uppala, and M. Fuentes, 2004: The ERA-40 archive. ERA-40 Project Report Series, No. 17, ECMWF, 31 pp.

Kiehl, J. T., J. J. Hack, G. B. Bonan, B. A. Boville, D. L. Williamson, and P. J. Rasch, 1998: The National Center for Atmospheric Research Community Climate Model: CCM3. *J. Climate,* **11,** 1151–1178.

Kistler, R., and Coauthors, 2001: The NCEP–NCAR 50-Year Reanalysis: Monthly means CD-ROM and documentation. *Bull. Amer. Meteor. Soc.,* **82,** 247–267.

Langematz, U., 2000: An estimate of the impact of observed ozone losses on stratospheric temperature. *Geophys. Res. Lett.,* **27,** 2077–2080.

Liang, X. Z., W. C. Wang, and J. S. Boyle, 1997: Atmospheric ozone climatology for use in general circulation models. Tech. Rep. 43, PCMDI, 25 pp.

Lin, S.-J., 2004: A "vertically Langrangian" finite-volume dynamical core for global models. *Mon. Wea. Rev.,* **132,** 2293–2307.

London, J., R. D. Bojkov, S. Oltmans, and J. I. Kelley, 1976: Atlas of the global distribution of total ozone. NCAR Tech. Note 113, 276 pp.

Manzini, E., and L. Bengtsson, 1996: Stratospheric climate and variability from a general circulation model and observations. *Climate Dyn.,* **12,** 615–639.

——, and N. A. McFarlane, 1998: The effect of varying the source spectrum of a gravity wave parameterization in a middle atmosphere general circulation model. *J. Geophys. Res.,* **103,** 31 523–31 539.

——, M. A. Giorgetta, M. Esch, L. Kornblueh, and E. Roeckner, 2006: The influence of sea surface temperature on the northern winter stratosphere: Ensemble simulations with the MAECHAM5. *J. Climate,* **19,** 3863–3881.

Newman, P. A., E. R. Nash, and J. E. Rosenfield, 2001: What controls the temperature of the Arctic stratosphere during the spring? *J. Geophys. Res.,* **106,** 19 999–20 010.

Pawson, S., and Coauthors, 2000: The GCM-Reality Intercomparison Project for SPARC: Scientific issues and initial results. *Bull. Amer. Meteor. Soc.,* **81,** 781–796.

Polvani, L. M., and D. W. Waugh, 2004: Upward wave activity flux as precursor to extreme stratospheric events and subsequent anomalous surface weather regimes. *J. Climate,* **17,** 3548–3554.

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

Rind, D., R. Suozzo, N. K. Balachandran, A. Lacis, and G. Russell, 1988: The GISS Global Climate–Middle Atmosphere Model. Part I: Model structure and climatology. *J. Atmos. Sci.,* **45,** 329–370.

——, D. Shindell, P. Lonergan, and N. K. Balachandran, 1998: Climate change and the middle atmosphere. Part III: The doubled CO_2 climate revisited. *J. Climate,* **11,** 876–894.

——, J. Lerner, K. Shah, and R. Suozzo, 1999: Use of on-line tracers as a diagnostic tool in general circulation model development: 2. Transport between the troposphere and stratosphere. *J. Geophys. Res.,* **104,** 9151–9167.

——, ——, J. Perlwitz, C. McLinden, and M. Prather, 2002: Sensitivity of tracer transports and stratospheric ozone to sea surface temperature patterns in the doubled CO_2 climate. *J. Geophys. Res.,* **107,** 4800, doi:10.1029/2002JD002483.

Roeckner, E., and Coauthors, 2003: The atmospheric general circulation model ECHAM5, Part I: Model description. Tech. Rep. 349, MPI, Hamburg, Germany, 140 pp.

Sassi, F., R. R. Garcia, B. A. Boville, and R. Roble, 2004: Effect of El Niño–Southern Oscillation on the dynamical, thermal, and chemical structure of the middle atmosphere. *J. Geophys. Res.,* **109,** D17108, doi:10.1029/2003JD004434.

Schmidt, G. A., and Coauthors, 2006: Present-day atmospheric simulations using GISS ModelE: Comparisons to in situ, satellite, and reanalysis data. *J. Climate,* **19,** 153–192.

Schnadt, C., and M. Dameris, 2003: Relationship between North Atlantic Oscillation changes and stratospheric ozone recovery in the Northern Hemisphere in a chemistry-climate model. *Geophys. Res. Lett.,* **30,** 1487, doi:10.1029/2003GL017006.

Shibata, K., H. Yoshimura, M. Ohizumi, M. Hosaka, and M. Sugi, 1999: A simulation of troposphere, stratosphere and mesosphere with an MRI/JMA98 GCM. *Papers Meteor. Geophys.,* **50,** 15–53

Shindell, D. T., D. Rind, and P. Lonergan, 1998: Increased polar stratospheric ozone losses and delayed eventual recovery owing to increased greenhouse gas concentrations. *Nature,* **392,** 589–592.

——, R. L. Miller, G. A. Schmidt, and L. Pandolfo, 1999: Simulation of recent northern winter climate trends by greenhouse gas forcing. *Nature,* **399,** 452–455.

Shine, K. P., and Coauthors, 2003: A comparison of model-simulated trends in stratospheric temperatures. *Quart. J. Roy. Meteor. Soc.,* **129,** 1565–1588.

Stolarski, R. S., A. R. Douglass, S. Steenrod, and S. Pawson, 2005: Trends in stratospheric ozone: Lessons learned from a 3D chemical transport model. *J. Atmos. Sci.,* **63,** 1028–1041.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 648 pp.