A process-based evaluation of biases in extratropical stratosphere-troposphere coupling in subseasonal forecast systems

Chaim I. Garfinkel[1], Zachary D. Lawrence[2,3], Amy H. Butler[4], Etienne Dunn-Sigouin[5], Irina Statnaia[6], Alexey Yu. Karpechko[6], Gerbrand Koren[7], Marta Abalos[8], Blanca Ayarzagüena[8], David Barriopedro[9], Natalia Calvo[8], Alvaro de la Cámara[8], Andrew Charlton-Perez[10], Judah Cohen[11], Daniela I.V. Domeisen[12,13], Javier García-Serrano[14], Neil P. Hindley[15], Martin Jucker[16], Hera Kim[17], Robert W. Lee[10], Simon H. Lee[18], Marisol Osman[19,20], Froila M. Palmeiro[14], Inna Polichtchouk[21], Jian Rao[22], Jadwiga H. Richter[23], Chen Schwartz[3], Seok-Woo Son[17], Masakazu Taguchi[24], Nicholas L. Tyrrell[6], Corwin J. Wright[15], and Rachel W.-Y. Wu[13]

[1]Fredy & Nadine Herrmann Institute of Earth Sciences, The Hebrew University of Jerusalem, Israel
[2]Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder, Boulder, CO, USA
[3]NOAA Physical Sciences Laboratory (PSL), Boulder, CO, USA
[4]NOAA Chemical Sciences Laboratory (CSL), Boulder, CO, USA
[5]NORCE Norwegian Research Centre and Bjerknes Centre for Climate Research, Bergen, Norway
[6]Finnish Meteorological Institute, Meteorological Research, Helsinki, Finland
[7]Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, The Netherlands
[8]Department of Earth Physics and Astrophysics, Universidad Complutense de Madrid, Madrid, Spain
[9]Instituto de Goeciencias (IGEO), Consejo Superior de Investigaciones Científicas – Universidad Complutense de Madrid (CSIC-UCM), Madrid, Spain.
[10]Department of Meteorology, University of Reading, Reading, UK
[11]Atmospheric and Environmental Research Inc., Lexington, MA, USA.
[12]University of Lausanne, Lausanne, Switzerland
[13]ETH Zurich, Zurich, Switzerland
[14]Group of Meteorology, Universitat de Barcelona (UB), Barcelona, Spain
[15]Centre for Climate Adaptation and Environment Research, University of Bath, Bath, UK
[16]Climate Change Research Centre and ARC Centre of Excellence for Climate Extremes, University of New South Wales, Sydney, Australia
[17]School of Earth and Environmental Sciences, Seoul National University, South Korea
[18]Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY
[19]CONICET – Universidad de Buenos Aires, Centro de Investigaciones del Mar y la Atmósfera (CIMA), Buenos Aires, Argentina
[20]CNRS – IRD – CONICET – UBA. Instituto Franco-Argentino para el Estudio del Clima y sus Impactos (IRL 3351 IFAECI), Buenos Aires, Argentina
[21]European Centre for Medium-Range Weather Forecasts, Reading, UK
[22]Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters / Key Laboratory of Meteorological Disaster of Ministry of Education, Nanjing University of Information Science and Technology, Nanjing, China
[23]Climate and Global Dynamics Laboratory, National Center for Atmospheric Research, Boulder, CO
[24]Department of Earth Science, Aichi University of Education, Kariya, Japan

Correspondence: Chaim I. Garfinkel (chaim.garfinkel@mail.huji.ac.il)

Abstract. Two-way coupling between the stratosphere and troposphere is recognized as an important source of subseasonal-to-seasonal (S2S) predictability and can open windows of opportunity for improved forecasts. Model biases can, however, lead

to a poor representation of such coupling processes; drifts in a model's circulation related to model biases, resolution, and parameterizations have the potential to feed back on the circulation and affect stratosphere-troposphere coupling. We introduce a set of diagnostics using readily available data that can be used to reveal these biases, and then apply these diagnostics to 22 S2S forecast systems.

In the Northern Hemisphere, nearly all S2S forecast systems underestimate the strength of the observed upward coupling from the troposphere to the stratosphere, downward coupling within the stratosphere, and the persistence of lower stratospheric temperature anomalies. While downward coupling from the lower stratosphere to the near surface is well represented in the multi-model ensemble mean, there is substantial inter-model spread likely related to how well each model represents tropospheric stationary waves.

In the Southern Hemisphere, the stratospheric vortex is over-sensitive to upward propagating wave flux in the forecast systems. Forecast systems generally overestimate the strength of downward coupling from the lower stratosphere to the troposphere, even as most underestimate the radiative persistence in the lower stratosphere. In both hemispheres, models with higher lids and a better representation of tropospheric quasi-stationary waves generally perform better at simulating these coupling processes.

1   Introduction

The extratropical stratosphere and troposphere are coupled through dynamical interactions between planetary-scale atmospheric Rossby waves and the mean flow. This vertical coupling operates in both directions: upward coupling from tropospheric variability induces variability in the stratosphere, while downward coupling from stratospheric variability can impact weather in the troposphere (Butler et al., 2019; Scaife et al., 2022). Both weak and strong polar stratospheric vortex extremes have been shown to influence surface climate and weather extremes for weeks to months afterwards (Domeisen and Butler, 2020) due to the long radiative timescales in the lower stratosphere (Hitchcock et al., 2013), which means that stratospheric variability can potentially provide windows of opportunity for prediction on subseasonal-to-seasonal (S2S) timescales (Butler et al., 2019; Domeisen et al., 2020b). However, model biases in either the troposphere or stratosphere can impact these coupling processes, compromising opportunities to increase S2S predictability that could otherwise be achieved. The goal of this study is to identify systematic biases in extratropical stratosphere-troposphere coupling processes across S2S forecast systems.

Variability in the upward flux of planetary-scale (wavenumbers 1-3) Rossby waves drives variability in the stratospheric polar vortex. Upward wave propagation is strengthened when the wave (or eddy) constructively interferes with the climatological stationary wave pattern, while weakened wave flux occurs when the linear interference is destructive (Garfinkel et al., 2010; Smith and Kushner, 2012). In addition, Rossby waves can amplify or weaken due to nonlinear processes (Scinocca and Haynes, 1998; Boljka and Birner, 2020). Rossby waves can only propagate upward into the stratosphere when the zonal flow is westerly but below a critical wind speed (Charney and Drazin, 1961), conditions that occur primarily in Northern Hemisphere (NH) extended winter (November-March) and Southern Hemisphere (SH) spring (September-November). A weaker upward flux of wave activity can lead to a strengthening of the polar vortex (Limpasuvan et al., 2004). On the other hand, an anomalously

strong or persistent pulse of wave activity can weaken, and even reverse, the westerly winds of the vortex (Andrews et al., 1987; Polvani and Waugh, 2004; Garfinkel et al., 2010). In particular, about once every two years the Arctic polar vortex completely breaks down and the zonal winds reverse direction in an extreme event called a "sudden stratospheric warming" (Baldwin et al., 2021). In SH spring, this upward coupling more typically manifests as a modulation of the timing of the seasonal polar vortex breakdown, with weaker upward flux of wave activity resulting in a delayed breakdown in spring, and vice versa for stronger upward wave flux (Byrne and Shepherd, 2018; Lim et al., 2018). A complete breakdown of the SH vortex has only been observed once, in September 2002.

Variability in the strength and location of the stratospheric polar vortex can also exert a downward influence on weather patterns (Boville, 1984; Haynes et al., 1991; Hitchcock and Simpson, 2014). Near the tropopause, interactions of the stratospheric signal with both transient and stationary eddies are important for communicating the signal to the surface (Song and Robinson, 2004; Domeisen et al., 2013; White et al., 2020, 2022). While both stratospheric and tropospheric factors influence the downward communication of the signal (Afargan-Gerstman et al., 2022), the exact mechanism of downward coupling remains unclear.

Accurately simulating both upward and downward vertical coupling requires reasonably accurate simulation of processes such as the location and strength of stationary planetary waves and the jet in the troposphere (Schwartz et al., 2022), the strength and seasonality of stratospheric wind speeds, and the radiative timescales of the lower stratosphere. Recently, Lawrence et al. (2022) identified systematic stratospheric biases across S2S forecast systems. In particular, they found that most forecast systems exhibit a warm bias in the global-mean stratosphere, and a cold bias in the extratropical lower stratosphere-upper troposphere. These biases were suggested to be due to biases in radiative heating rates associated with model biases in ozone and water vapor (cf. Bland et al., 2021). Most forecast systems also showed strong and cold polar vortex biases, which suggests that there are underlying difficulties in accurately representing vertical coupling processes. In general, stratospheric biases were substantially worse for models with a low model lid height, a longstanding issue (Lawrence, 1997; Marshall and Scaife, 2010) that has also been identified in seasonal prediction systems (Butler et al., 2016) and climate models (Charlton-Perez et al., 2013), which can be exacerbated by poorly designed physics parameterizations (Shaw and Perlwitz, 2010).

While systematic biases in the stratosphere were detailed in Lawrence et al. (2022), a deeper exploration of how S2S models simulate the processes that underlie stratosphere-troposphere vertical coupling is warranted, given that these processes ultimately drive the impacts on surface weather patterns and regional hazards. As part of the collaborative effort of the World Climate Research Programme (WCRP) Atmospheric Processes and their Role in Climate (APARC) Stratospheric Network for the Assessment of Predictability (SNAP) activity, we investigate how extratropical atmospheric biases are linked to the simulation of stratosphere-troposphere coupling in S2S forecast systems. After introducing the data and methods in section 2, we demonstrate that many S2S forecast systems struggle to represent the strength of observed upward coupling from the troposphere to the stratosphere (Section 3.1), the sensitivity of the stratospheric polar vortex to upward wave flux (Section 3.2), interannual variability in heat flux extremes (Section 3.3), downward coupling within the stratosphere (Section 3.4), and downward coupling from the lower stratosphere to the surface (Section 3.5). After considering possible factors that can account

for the intermodel spread in coupling strength (Section 3.6), we summarize our results and place them in the context of previous work (Section 4).

2 Data & Methods

2.1 Subseasonal-to-Seasonal (S2S) Hindcast and Reanalysis Datasets

We use ensemble hindcast data from the S2S Prediction Project Database (Vitart et al., 2017), and depending on data availability, select forecast systems not included in the S2S database: (i) the National Oceanic and Atmospheric Administration's Global Ensemble Forecast System version 12 (NOAA GEFSv12; Hamill et al., 2021; Guan et al., 2021), (ii) the National Center for Atmospheric Research Community Earth System Model version 2 (CESM2) with version 6 of the Community Atmosphere Model as its atmospheric component (NCAR CESM2-CAM6, hereafter CESM2-CAM), and (iii) CESM2 with the version 6 of the Whole Atmosphere Community Climate Model as its atmospheric component (CESM2-WACCM6, hereafter CESM2-WACCM; Richter et al., 2022). Daily gridded latitude-longitude data was only retained for the seven forecast systems that provide at least 35-day forecasts to the S2S database due to the large data volume, and so metrics which rely on this data are only computed for these seven systems.

Lawrence et al. (2022) analyzed biases over the period common to all models (1999-2010), but here we include upgraded versions of several models, for which the hindcasts begin several years after 1999. Furthermore, the specific days on which forecasts are initialized differ across systems even for a given year. We therefore have elected not to focus on a common period in this paper for all analyses except those in Section 3.3. The specific model versions and the period used for each system are included in Table 1, and their vertical resolutions are detailed in Figure 1. For "pixel figures" quantifying biases in individual systems (e.g., Figure 2, 4), we subsample reanalysis data to match each system, thus allowing us to pinpoint biases. For figures showing lagged correlations and lagged regression, we show the mean across the forecasting systems of the subsampled coupling strength with a solid black line, and the spread in the subsampled coupling strength across the available S2S systems with a vertical thin line; because there is no exact overlap in the analysis period, model biases should not be inferred from face value from these lagged correlation/regression figures. Nonetheless, these thin vertical lines offer an estimate of the range of sampling variability in ERA5, and thus if a given model lies outside of this range, a bias can be even more confidently detected as internal variability is no longer a reasonable explanation.

The subseasonal hindcasts analysed here are initialized with different atmospheric datasets. To ensure this has no significant effects on our results, we compare the hindcast fields to those from the ERA5 reanalysis (Hersbach et al., 2020) so that comparisons and biases are all determined with respect to a consistent dataset. Note that for the time periods and levels considered here (post-1990 and up to 10-hPa) most modern reanalysis products are in good agreement (Long et al., 2017; Gerber and Martineau, 2018; Ayarzagüena et al., 2019; Fujiwara, M. et al., 2021), and thus our results should be robust across reanalyses.

Table 1. Details of the subseasonal-to-seasonal forecast systems used herein.

Model	S2S Database Version(s)	Hindcast Period	Inits per season	Ensembles	Forecast Span	Model Top
BoM*	POAMA P24	1999-2010	15	33	62 days	10 hPa
CESM2-CAM*	–	1999-2019	12-13	11	45 days	2 hPa
CESM2-WACCM	–	1999-2019	12-13	5	45 days	4.5×10^{-6} hPa
CMA*	BCC-CPS-S2Sv1	1994-2014	90-91	4	60 days	0.5 hPa
CMA	BCC-CPS-S2Sv2	2005-2019	25	4	60 days	0.1 hPa
CNR-ISAC*	GLOBO	1990-2010	18	5	32 days	6.8 hPa
CNRM	CNRM-CM 6.0	1993-2014	12	15	50 days	0.01 hPa
CNRM	CNRM-CM 6.1	1992-2017	13	10	47 days	0.01 hPa
ECCC-lo*	GEPS 4	1995-2014	25-26	4	32 days	2 hPa
ECCC-hi	GEPS 6	1998-2017	23-25	4	32 days	0.1 hPa
ECCC-hi	GEPS 7	2001-2020	13-21	4	32 days	0.1 hPa
ECMWF	CY45R1	1998-2018	26	11	46 days	0.01 hPa
ECMWF	CY47R3	2002-2020	26	11	46 days	0.01 hPa
GEFSv12	–	2000-2019	12-13	11	35 days	0.1 hPa
HMCR	RUMS	1991-2014	12-13	11	46 days	0.04 hPa
JMA	GEPS1701	1990-2012	9	5	34 days	0.01 hPa
JMA	CPS3	1991-2020	6	5	34 days	0.01 hPa
KMA	GloSea5-GC2	1991-2016	12	3	60 days	85 km
KMA	GloSea6-GC32	1993-2016	12	3	60 days	85 km
NCEP	CFSv2	1999-2010	90-91	4	44 days	0.02 hPa
UKMO	GloSea5	1993-2016	12	7	60 days	85 km
UKMO	GloSea6	1993-2016	12	7	60 days	85 km

* Systems with low-top models. Note that we use the high-top HMCR RUMS model version.

2.2 Methods

We use the following eight key metrics to diagnose coupling strength throughout this paper: $\overline{v_{k=1}T_{k=1}}$ at 500 hPa and 100 hPa; $\overline{v_{k=2}T_{k=2}}$ at 500 hPa and 100 hPa; polar cap height (60°-pole) at 10 hPa, 100 hPa, and 850 hPa; and polar cap temperature (60°-pole) at 100 hPa. v denotes the meridional wind, T the temperature, and ⁻ the zonal mean. We decompose v and T by wavenumber before computing their product, e.g., $\overline{v_{k=1}T_{k=1}}$ for zonal wavenumber-1 (wave-1) heat flux. The upward flux of planetary waves is diagnosed using the meridional eddy heat flux, (e.g., $\overline{v_{k=1}T_{k=1}}$) rather than the vertical component of the Eliassen-Palm flux due to the limited vertical resolution available in the S2S archive.

S2S models typically archive data at coarse resolution due to the huge data volume. For models on the S2S database, we consider instantaneous daily values at 00 UTC on a 1.5° × 1.5° latitude/longitude grid, with 10 pressure levels between 1000
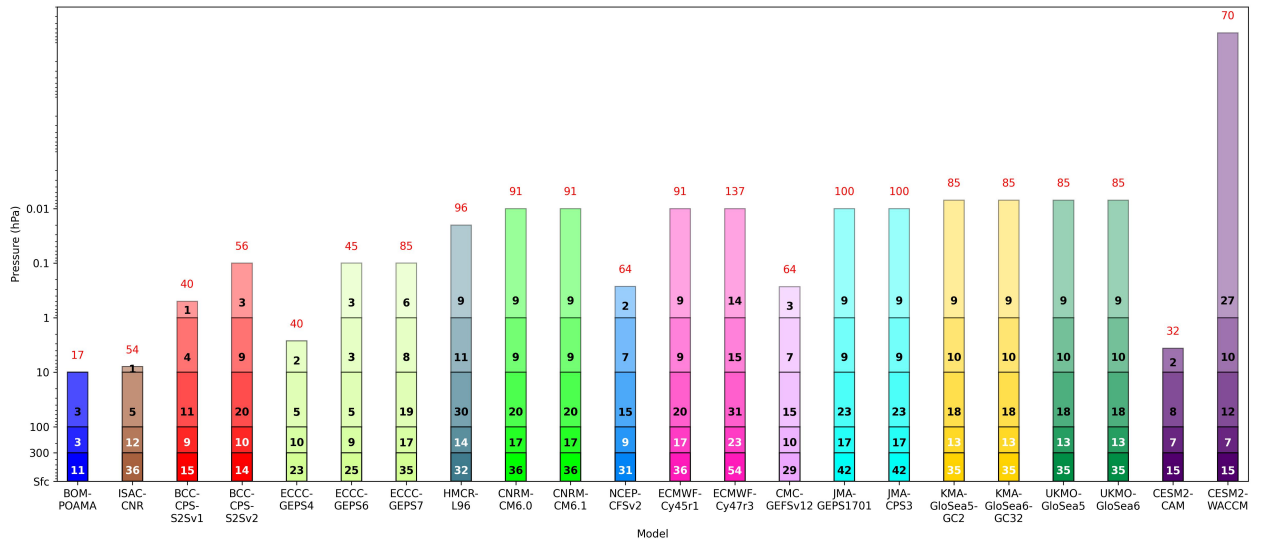
Figure 1. Schematic representation of model vertical resolution for all S2S prediction systems used in this study. Each block represents the pressure range indicated on the y axis. The number of model levels in each range is shown numerically (the font color was chosen for visualization and does not have additional meaning). The red number at the top of each bar shows the total number of levels in each model setup.

and 10 hPa. GEFSv12 data are provided 6-hourly on a $0.5° \times 0.5°$ grid, with 25 pressure levels respectively between 1000 and 1 hPa. CESM2-CAM and CESM2-WACCM provide zonally averaged daily fields at 192 latitudes ($\sim0.9424°$ resolution) on the pressure levels closest to the model levels, which we interpolate to a set of 32 standard pressure levels between 1000 and 10 hPa. Heat flux data is not available for CESM2-CAM, and hence we show this model for limited diagnostics only. The eight diagnostics are computed on the available model grid.

As in Lawrence et al. (2022), we define forecast systems with model tops at or above 0.1 hPa with several levels above 1 hPa as "high-top" models, and all others as "low-top" models. Using this definition results in 17 forecast systems with high-top models and 5 forecast systems with low-top models (see Table 1); however, not all models are included for each analysis. Low-top models are identified with asterisks and/or dotted lines in the figures. We stress that the computation of high-top and low-top means is obtained from an unbalanced distribution of high-top and low-top models.

For each variable and forecast system, we derive lead-time-dependent climatologies, which we subtract from the raw forecast quantities to determine forecast anomalies. These climatologies are calculated by averaging all ensemble-mean hindcasts for a given day of year and for each lead time. For those systems providing a fixed set of hindcast initializations that do not uniformly

cover the same days of year in the hindcasts (e.g., GEFSv12 and CESM2), we permit differences of up to three days when creating the lead-time-dependent climatologies.

125   We quantify the tightness of coupling by using both regression and correlation analyses. Regression coefficients directly diagnose the strength of coupling, and are the closest we can get to answering questions such as "what is the heat flux anomaly at 100-hPa for a given heat flux anomaly at 500-hPa?". The downside of regression is that it is not possible to meaningfully compare the different coupling metrics in the paper to see their relative importance because the units are different. On the other hand, correlations normalize the units and allow for comparison between different metrics. Correlations also quantify

130 how much of the linear variability between two quantities is shared. For most models and metrics, regression and correlation coefficients are similar. However, there are notable exceptions if a given model fails to simulate a reasonable amount of variance for a given metric. In these cases, we elect to use regression to diagnose coupling strength, as the correlation conflates two possible sources of error: error in the coupling strength with error in the underlying variance (the Results section provides several examples of such behavior). For completeness, we present in the supplemental material figures diagnosing coupling

135 strength using correlation. In all cases we calculate the regression and correlation coefficients for individual ensemble members first and then average over members.

    Some models suffer from large (>40%) biases in variance, and so this concern about variance biases complicating the interpretation of correlations is difficult to sidestep. We demonstrate this in Figure 2, which shows the percentage error in the daily variance in each forecast system for our eight key metrics and days 22-28 of the forecast. We compare each forecast

140 system to the corresponding period in ERA5, and if all available ensemble members show a bias of the same sign, we indicate that pixel with a "×" symbol. Applying an f-test leads to a larger proportion of the pixels indicating significant biases (not shown). Most models overestimate variance in lower tropospheric polar cap height and tropospheric planetary wave heat flux in both hemispheres. In contrast, most underestimate variance in the lower stratosphere in the SH, and in the NH there is a notable decrease in the magnitude of the bias in variance from the troposphere up to the stratosphere. Most models also

145 suffer from too little variance in lower stratospheric polar cap height and temperature in the SH. These biases in variance are qualitatively similar though weaker earlier in the integration (e.g., days 8-14, not shown). We are not aware of previous work that has found such too-strong variance biases in the troposphere, and the causes and implications of these biases should be explored in future work.

3   Results

150 We now consider the relative abilities of the forecast systems to capture the physical processes underlying stratosphere-troposphere coupling. To do this, we subdivide stratosphere-troposphere coupling into several components as follows and consider each individually below:

   1. Vertical propagation of planetary waves from the troposphere into the stratosphere

   2. The sensitivity of the stratospheric polar vortex to upward wave driving from the lowermost stratosphere
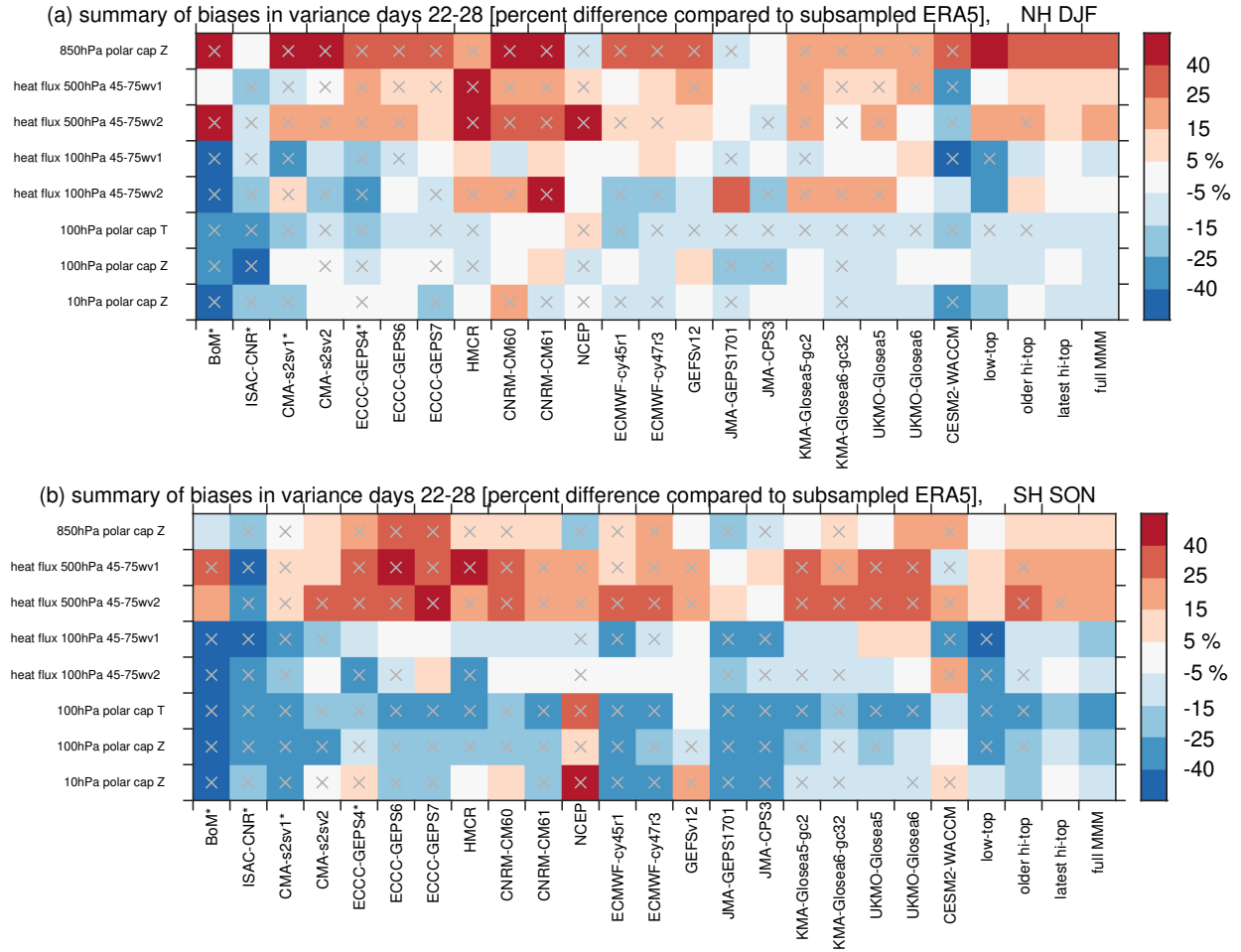
Figure 2. Variance of daily values of the various diagnostics in days 22-28 in the (a) Northern Hemisphere and (b) Southern Hemisphere. For each forecast system we compare the variance to that for the corresponding period in ERA5, and then show the percent error. A gray '×' indicates models and metrics for which all ensemble members simulate a variance that is either too-weak or too-strong, or alternatively if the ERA5 variance doesn't fall within the envelope of the available members. The bias is defined as the difference in variance between the model and ERA5 divided by the variance in ERA5. The last four columns show the mean bias for low-top models, the older versions of high-top models, the latest versions of high-top models, and all models; A gray '×' indicates all models agree on the sign of the bias.

3. Sufficient interannual spread in daily heat flux extremes

4. Downward propagation of stratospheric polar vortex anomalies from the upper- and mid- stratosphere to the lower stratosphere related to wave-mean flow interactions

5. The persistence of the polar vortex signal in the lower stratosphere that arises due to the long radiative timescales

6. Downward propagation from the lower stratosphere to the near-surface.

3.1 Vertical propagation of planetary waves from the troposphere into the stratosphere

We begin by considering the upward coupling of wave activity from the troposphere to the stratosphere. This coupling is quantified by computing the lagged correlation and regression coefficients between 500-hPa and 100-hPa heat fluxes averaged over 45-75° in each hemisphere (Figure 3ab). The dominant direction of coupling is for tropospheric (500-hPa) heat fluxes to precede lower stratospheric (100-hPa) heat fluxes.

In the NH, this coupling peaks when tropospheric heat flux precedes lower stratospheric heat flux by 3 days for wave-1 and by 2 days for wave-2 in ERA5 (thick black lines in Figure 3). At this lag, a 1 Km/s anomaly at 500hPa is associated with a 1.91 Km/s anomaly at 100hPa, with a correlation of 0.46. While the forecast systems capture this behavior qualitatively, most underestimate the magnitude of the correlation and regression for wave-1. For the high-top models, biases identified in the regression coefficients are mirrored by the biases in correlations. However, for the low-top models this is not the case. For example, BoM has one of the highest correlations for wave-1 of any model, while its regression coefficient is the lowest. This is due to the fact that BoM underestimates the variance of wave-1 100-hPa heat fluxes by more than 50% (Figure 2a), and this underestimation is likely a reflection of the model's poor simulation of stratospheric variability more generally as documented in Domeisen et al. (2020a). While most high-top models do not show strong biases in the wave-1 100-hPa heat flux variance (Figure 2a), this may be a case of two biases canceling each other, i.e. too much tropospheric wave-1 variability being compensated for by too little upward wave propagation.

The bias in the regression coefficient for each model when compared to the corresponding period in ERA5 is shown in the top row of Figure 4. Most models suffer from too-weak upward coupling, with only two models (NCEP and JMA-CPS3) simulating a stronger regression coefficient than ERA5. The multi-model mean regression coefficient is low-biased by 11% for high-top models, and by 25% for low-top models. Figure 5a further considers the relationship between model biases in coupling strength and vertical lid height by contrasting the bias in coupling with the base-10 logarithm of the vertical lid pressure. While the model lid level is anticorrelated with wave-1 upward coupling strength ($r = -0.34$, not significant), a more pronounced and statistically significant effect is evident when comparing coupling strength with the magnitude of the climatological wave-1 heat flux in the troposphere (an indication of how well each model represents quasi-stationary waves, $r = 0.5$; Figure 6a). Models with a better representation of the climatological quasi-stationary wave-1 better represent its upward coupling. This effect is even more pronounced if we compare climatological heat flux at 100-hPa with the coupling strength ($r = 0.70$; not shown).

The upgrade of the CMA system from low-top to high-top led to a 29% reduction in its bias in wave-1 upward coupling strength, while the transition from ECCC-GEPS4 to ECCC-GEPS7 led to a 67% reduction in its bias (Figure 4a). Of the high-top models, CNRM struggles the most with the upward coupling strength, and the upgrade from CNRM-CM6.0 to CNRM-CM6.1 improved the fidelity of the simulation by 21%. GloSea6 (both KMA and UKMO) improved by 47% over GloSea5. ECMWF-cy47r3 is also improved over its earlier version, though the earlier version was already among the most realistic across all forecast systems and hence there was less room for improvement.

This overall underestimate of wave-1 upward coupling is confirmed in Figure 7, which shows the regression coefficient between 500-hPa height anomalies and the wave-1 heat flux at 100-hPa 3 days afterwards for December and January initial-izations. This analysis is performed for only seven of the models due to data availability and storage constraints. Consistent with previous work (e.g., Garfinkel et al., 2010), low heights in the Northwest Pacific and high heights in the Atlantic sector are associated with pulses of wave-1 heat flux in the lower stratosphere. These anomalies are in phase with the climatological wave-1, and thus constructively interfere with it. The models systematically underestimate the regression coefficient in both sectors. The low-top CMA and BoM are particularly biased, again revealing the importance of the model top.

The above results suggest that the S2S forecast systems need a higher model lid and more realistic stationary waves in the troposphere to simulate realistic upward wave-1 coupling between 500 and 100-hPa in the boreal winter. Biases are smaller for wave-2 upward coupling in the NH winter. Coupling is too strong in 13 of 21 models (Figure 4a), and the multi-model mean bias is 2.5% too strong. JMA-GEPS1701 simulates a coupling strength 49% stronger than in ERA5, however in its updated version (JMA-CP3) the bias drops to 7.9%. The mean bias of the regression coefficient is larger for low-top versus high-top models: specifically, coupling is 9% too-weak in low-top vs. 1.6% too-strong in the most recent version of high-top models.

BoM suffers from an unrealistically strong correlation (Figure 3). However, its upward coupling regression coefficient is the weakest among all models with a too-weak bias of 24%. This apparent paradox is, as before, due to its too-weak wave-2 variability at 100-hPa. The wave-2 coupling strength is significantly correlated to the model lid ($r = -0.49$, Figure 5b), and to the climatological stationary wave-2 in the lower stratosphere ($r = 0.45$ for climatological $\overline{v'T'}$ at 100-hPa), but not in the troposphere. Finally, the forecast systems better capture the tropospheric precursors of 100-hPa wave-2 heat flux as compared to wave-1 heat flux, with CNRM and UKMO in particular simulating regression coefficients of reasonable magnitude (Supplemental Figure S1).

In the Southern Hemisphere spring, models have systematically too-strong variance in tropospheric (500-hPa) planetary wave heat flux, with the exception of ISAC-CNR which underestimates the variance (Figure 2b). In contrast, lower stratospheric (100-hPa) planetary wave heat flux is generally too weak in most models. The multi-model mean regression between the 500-hPa and 100h-Pa wave-1 heat flux is 6% too strong (Figure 3c,d), however there is a large spread among the models (Figure 4b). High-top models overall perform better: the model lid and regression coefficients are significantly correlated ($r = -0.56$ for both wave-1 and wave-2; Figure 5g,h). Biases are also smaller in models with better climatological stationary waves (Figure 6g,h), though this relationship is sensitive to the inclusion of BoM.
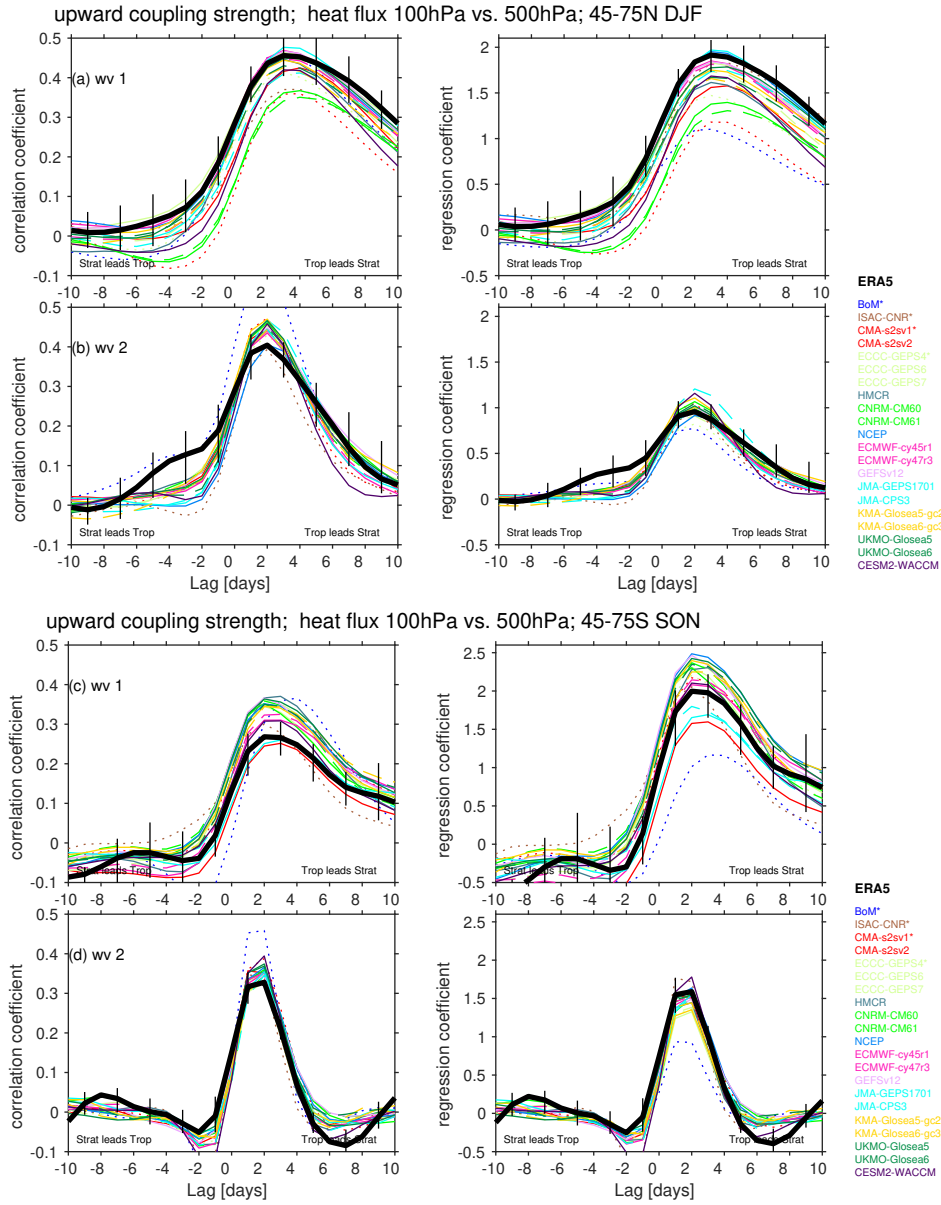
Figure 3. Coupling of $\overline{v'T'}$ 45-75° at 500hPa with that at 100hPa measured in terms of correlation coefficient (left column) and regression coefficient (right column). $\overline{v'T'}$ at 500hPa is taken from days 11 to 22, and we range $\overline{v'T'}$ at 100hPa from 10 days prior (i.e. days 1 to 12) to 10 days after (i.e. days 21 to 32). Low top models are dotted. Older versions of high-top models are dashed. Black vertical lines shows the range in coupling strength in ERA5 upon subsampling to match each of the 21 S2S forecast systems, and the solid black line indicates the mean of these 21 coupling strengths from ERA5. Panels a and b are for the NH and December-January-February, while panels c and d are for the SH and September-October-November. Panels a and c correspond to wave-1 (wv1) and panels b and d correspond to wave-2 (wv2).
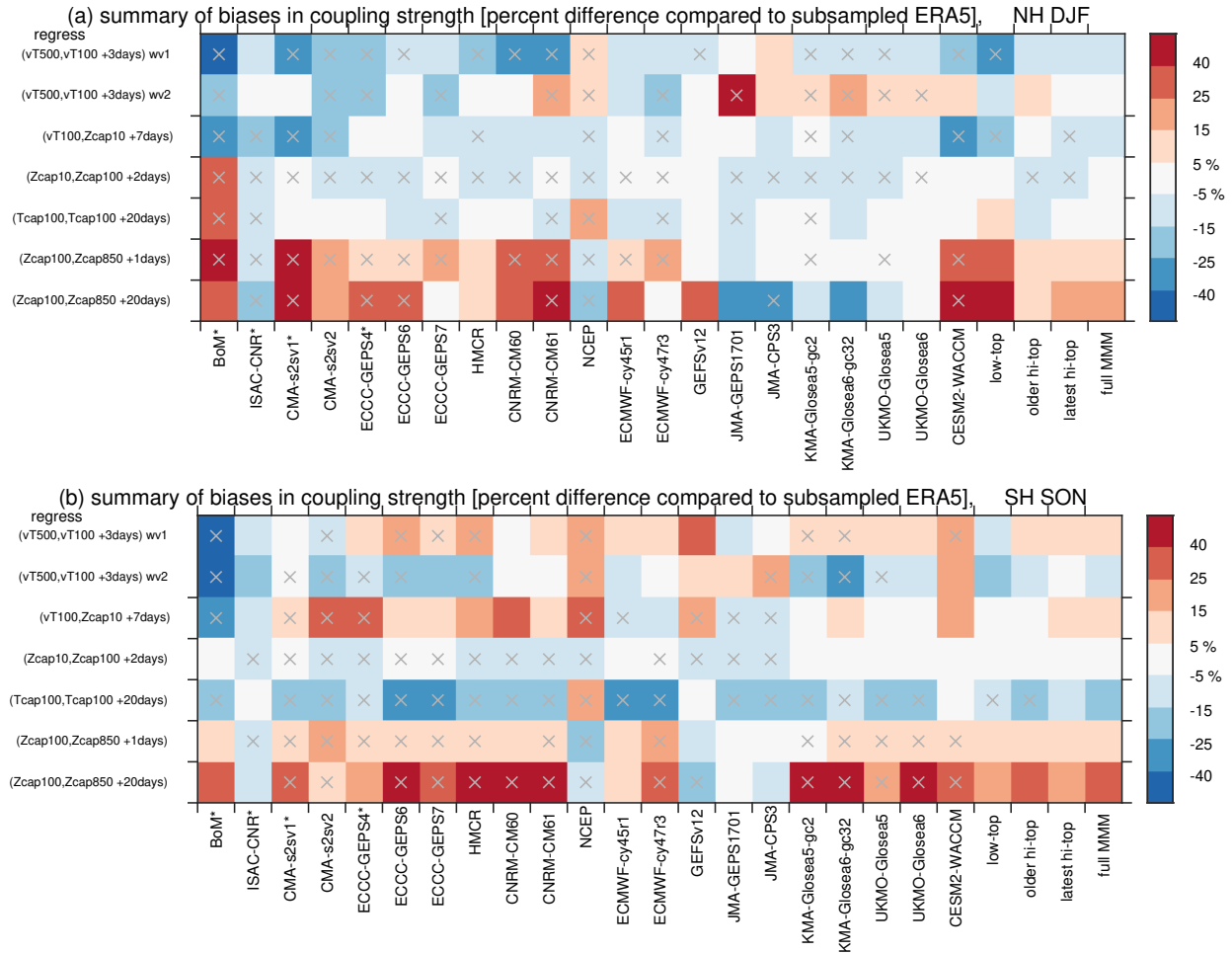
Figure 4. Summary of the biases in coupling strength. For each forecast system we compare to the coupling strength for the corresponding identical period in ERA5, and then show the percentage error. The bias is defined as the difference in coupling strength between the model and ERA5 divided by the coupling strength in ERA5 for the corresponding dates of each model. A gray '×' indicates models and metrics for which all ensemble members simulate a bias in the coupling strength of the same sign, or alternatively if ERA5 doesn't fall within the envelope of the available members. Low-top models are denoted with an asterisk after their name. Coupling strength is defined using regression, and the analogous figure for correlation is shown in Supplemental Figure S2. (top row) Upward coupling between $\overline{v'T'}$ wave-1 at 500hPa and at 100hPa lagged 3 days (cf. Figure 3); (second row) as in first row but for wave-2; (third row) Sensitivity of Z10hPa polar cap to 100hPa heat flux lagged by 7 days (cf. Figure 8); (fourth row) Coupling strength of Z10hPa polar cap with Z100hPa polar cap lagged 2 days (cf. Figure 12); (fifth row) Persistence of T100hPa polar cap on day 20 (cf. Figure 14); (sixth, seventh rows) Coupling strength of Z100hPa polar cap with Z850hPa polar cap lagged 1 day and 20 days (cf. Figure 13).
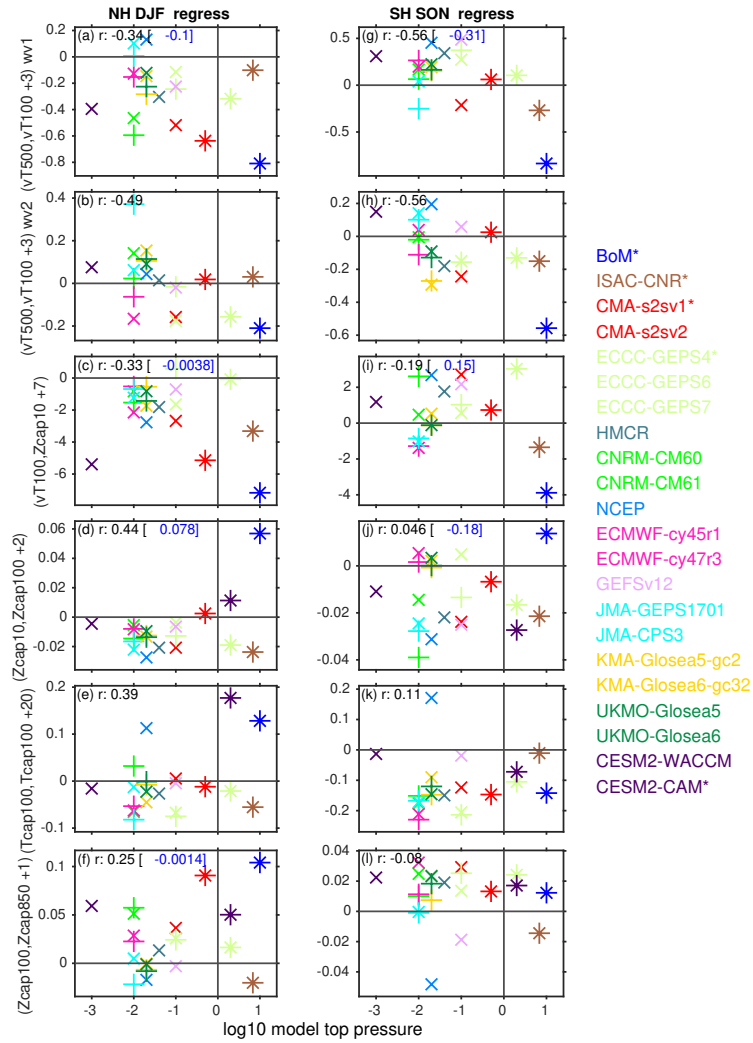
Figure 5. Relationship between (y-axis) the bias in coupling strength as compared to ERA5 for corresponding dates and (x-axis) the logarithm of the vertical lid for each model. Coupling strength is defined using regression; the corresponding figure for correlations is in Supplemental Figure S3. The correlation for each panel is indicated, and also the correlation without BoM in blue if this correlation differs from the overall correlation by more than 0.2. The left column is for the NH in DJF, and the right column is for the SH in SON. Low top models are indicated by an '*', older versions of high-top models with a '+', and the latest version of high-top models with an '×'. (a,g) upward wave-1 coupling on day 3 from Figure 3; (b,h) upward wave-2 coupling on day 3 from Figure 3; (c,i) sensitivity of Z10hPa polar cap to 100hPa heat flux on day 7 from Figure 8; (d,f) coupling strength of Z10hPa polar cap to Z100hPa polar cap on day 2 from Figure 12; (e,k) persistence of T100hPa polar cap on day 20 from Figure 14; (f,j) coupling strength of Z100hPa polar cap to Z850hPa polar cap on day 1 from Figure 13. The lid of GloSea is at 85km; we represent this with a value of 0.02hPa. The lid of WACCM is at 140km; since the levels in the ionosphere are not expected to improve the representation of stratosphere-troposphere coupling, we represent this model with a lid at 0.001hPa (still the highest lid of any model). A null hypothesis of no relationship can be rejected at the 95% confidence level using a two-sided Student-t test for correlations exceeding 0.42.
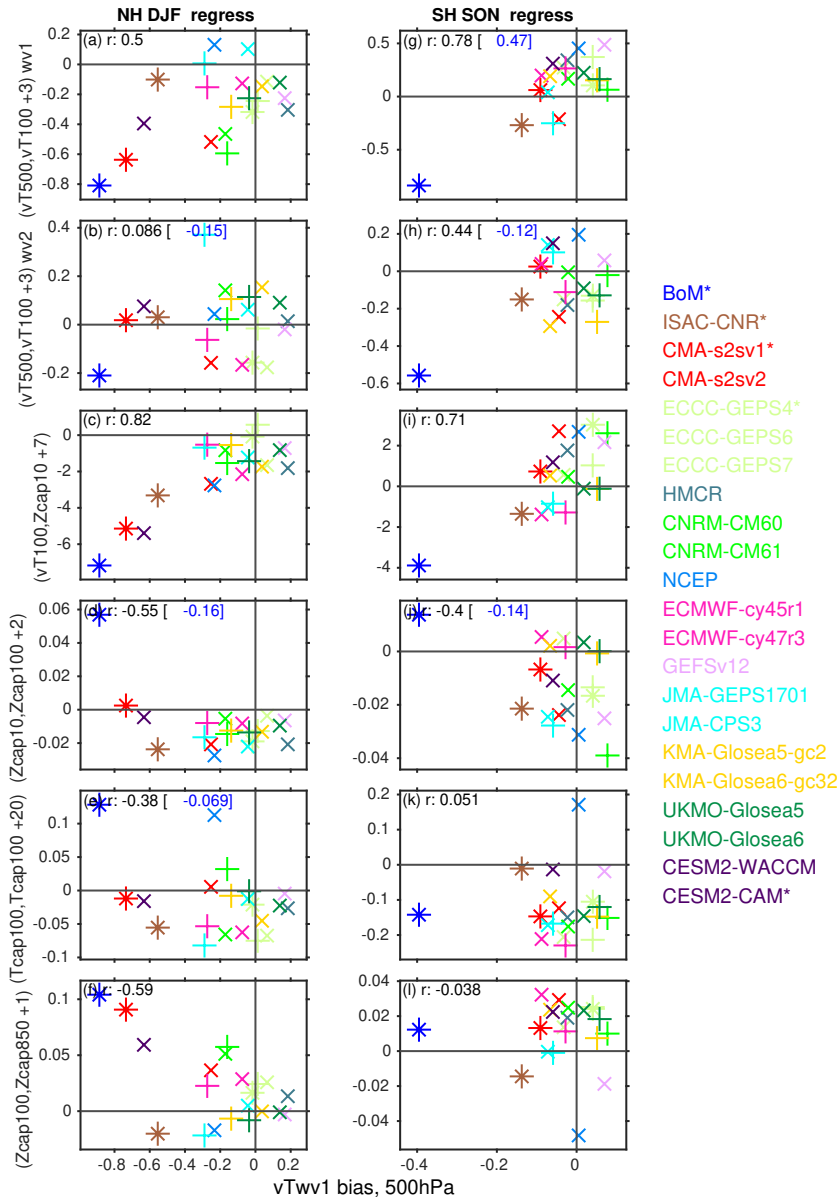
Figure 6. As in Figure 5, but for climatological wave-1 $\overline{v'T'}$ bias in days 22-28 at 500-hPa from 45-75° on the x-axis. The corresponding figure for correlations is in Supplemental Figure S4.

regress(Fz$_{100, k=1}$,Z500hPa) for DJ initializations, weeks: 3-4
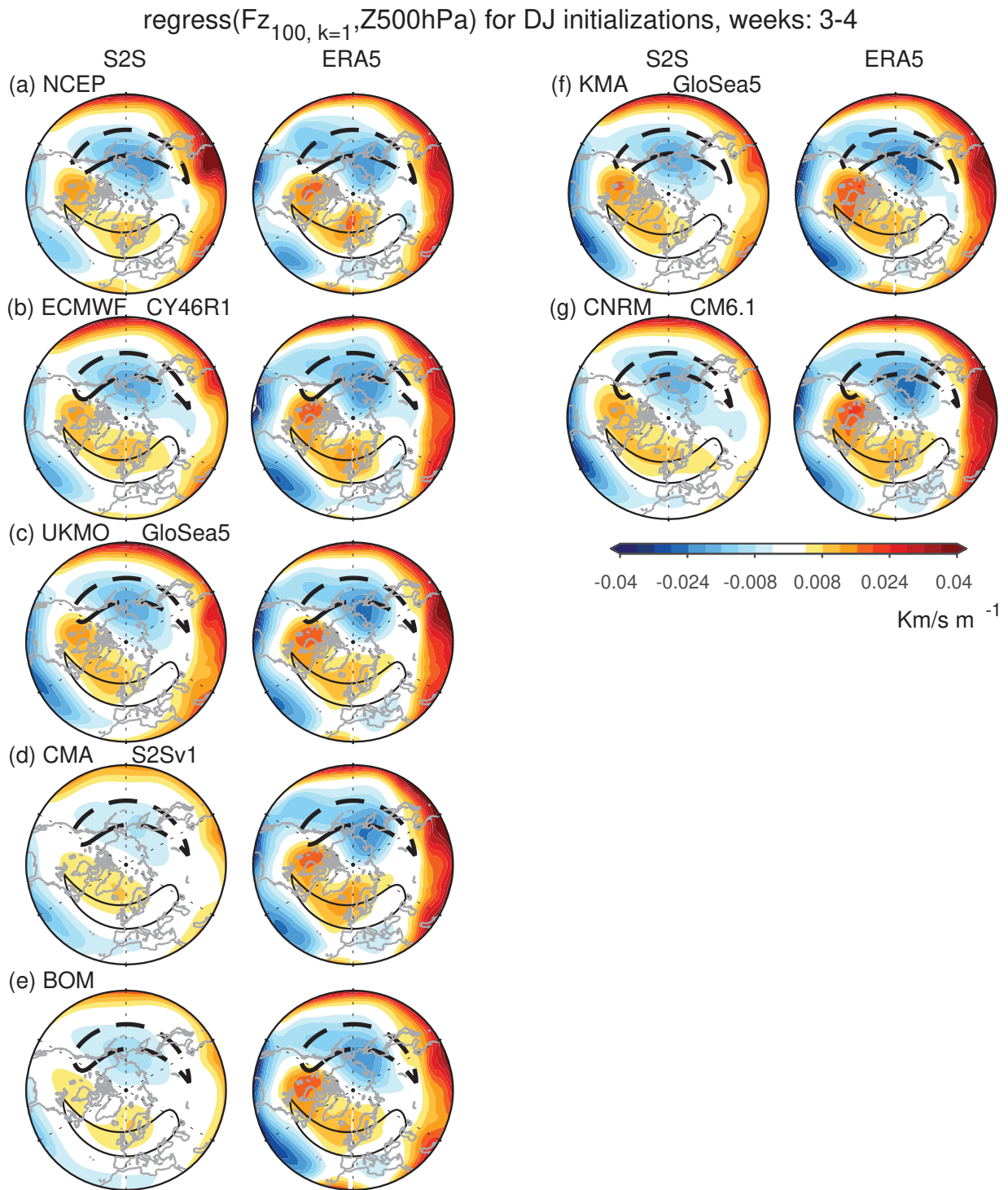
Figure 7. Maps of the regression coefficient between Z500 anomalies and $v'T'_{100hPa,k=1}$ anomalies two days later. Uses week 3 and 4 of December/January initializations. For each model we show the ERA5 subsampled to match each forecast system. The climatological wave-1 of Z500 between 45 and 75°N is shown with black contours.

3.2 The sensitivity of the stratospheric polar vortex to upward wave driving from the lowermost stratosphere

In order for models to fully capture the effect of tropospheric variability on the polar vortex, they must not only capture the upward flux of wave activity from the troposphere to the lower stratosphere, but also simulate a reasonable sensitivity of the polar vortex to lower stratospheric wave activity. We diagnose this sensitivity of the polar vortex by computing the lagged correlation and regression between 10-hPa polar cap height anomalies and the sum of wave-1 and wave-2 100-hPa heat flux (Figure 8).

In the Northern Hemisphere, the reanalysis correlation peaks when polar cap mid-stratospheric heights lag lower-stratospheric heat flux by 7 days, and most models simulate a similar lag (Figure 8a). Most models underestimate the magnitude of the coupling, however: the regression coefficient at lag 7-days is too weak in all models except ECCC-GEPS6, with the BoM, CMA and CESM2-WACCM models producing particularly large biases (Figure 4a). This underestimation is pronounced for the low-top models (too-weak bias of 23% in low-top vs. 9% in high-top). Models with a stronger bias in climatological 500-hPa heat flux suffer from a particularly pronounced too-weak sensitivity ($r = 0.82$; Figure 6c). Similarly, models with a cold-vortex bias also suffer from a too-weak sensitivity ($r = 0.53$, Supplemental Figure S5c). These effects are more important in explaining intermodel spread than the model lid (Figure 5c). The models are similarly biased if we contrast 100-hPa heat flux to the tendency of 10-hPa polar cap height (e.g., figure 7 of Dunn-Sigouin and Shaw (2015), not shown).

The net effect of the models' (i) underestimation of upward wave propagation from 500-hPa to 100-hPa and (ii) under-sensitive polar vortex to 100-hPa heat flux is that NH stratospheric polar variability is not coupled strongly enough to tropospheric variability. This is summarized in Figure 9, which shows maps of the regression coefficient between 500-hPa height anomalies and the tendency in 10-hPa polar cap heights over a ten-day period, analogous to figure 1 of Garfinkel et al. (2010). Low tropospheric heights in the North Pacific and high tropospheric heights over Scandinavia and the Ural mountains precede weakening of the vortex, but the regression coefficients are underestimated by all models. Note that NCEP is the least-biased model, and this model is the only one which overestimates upward coupling of 500-hPa heat flux with 100-hPa heat flux, although it still has an under-sensitive vortex to 100-hPa heat flux. The low-top BoM and CMA are the most biased in terms of upward coupling. UKMO and CNRM capture the effect of the Ural high on the vortex, but they underestimate that of the North Pacific low; recall that these models also succeed in simulating the tropospheric precursors of 100-hPa wave-2 heat flux (Supplemental Figure S1).

Finally, Figure 8a shows that there are negative correlation and regression coefficients between polar cap height and 100-hPa heat flux when polar cap height leads heat flux. In other words, a stronger polar vortex tends to precede weakened heat flux, while a weaker polar vortex tends to precede strengthened heat flux. This is associated with the polar vortex's ability to regulate its own wave-driving (Matsuno, 1970). The models estimate this effect accurately in the multi-model mean (bias less than 3%). The model which most strongly underestimates this effect is GloSea5 (both KMA and UKMO), however there is a marked improvement in GloSea6, with biases dropping from 22% to 8%. BoM and JMA-CP3, on the other hand, overestimate this effect. There is no relationship across models between this effect and either the model lid, the climatological wave-1 strength, or the cold pole bias.
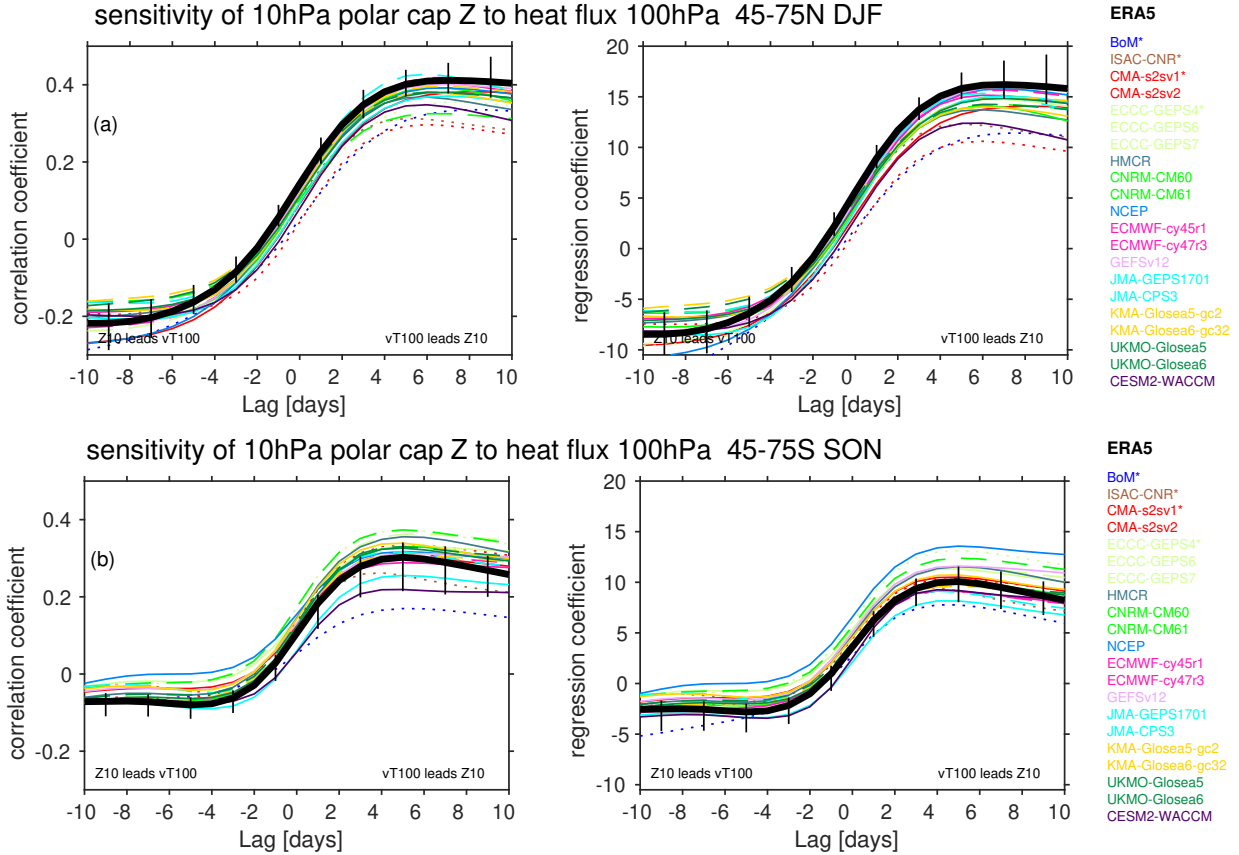
Figure 8. Sensitivity of 10-hPa polar cap Z to $\overline{v'T'}_{k=1+2}$ at 100-hPa 45-75°. $\overline{v'T'}_{k=1+2}$ at 100-hPa is taken from days 11 to 22, and we range 10hPa polar cap Z from 10 days prior to 10 days after. Low top models are dotted. Older versions of high-top models are dashed. The SH $\overline{v'T'}$ is multiplied by -1 before the analysis is performed to allow for a simpler comparison to the panels for the NH. Black vertical lines shows the range in coupling strength upon subsampling ERA5 reanalysis to match each of the forecast systems, and the solid black line indicates the mean of these coupling strengths.
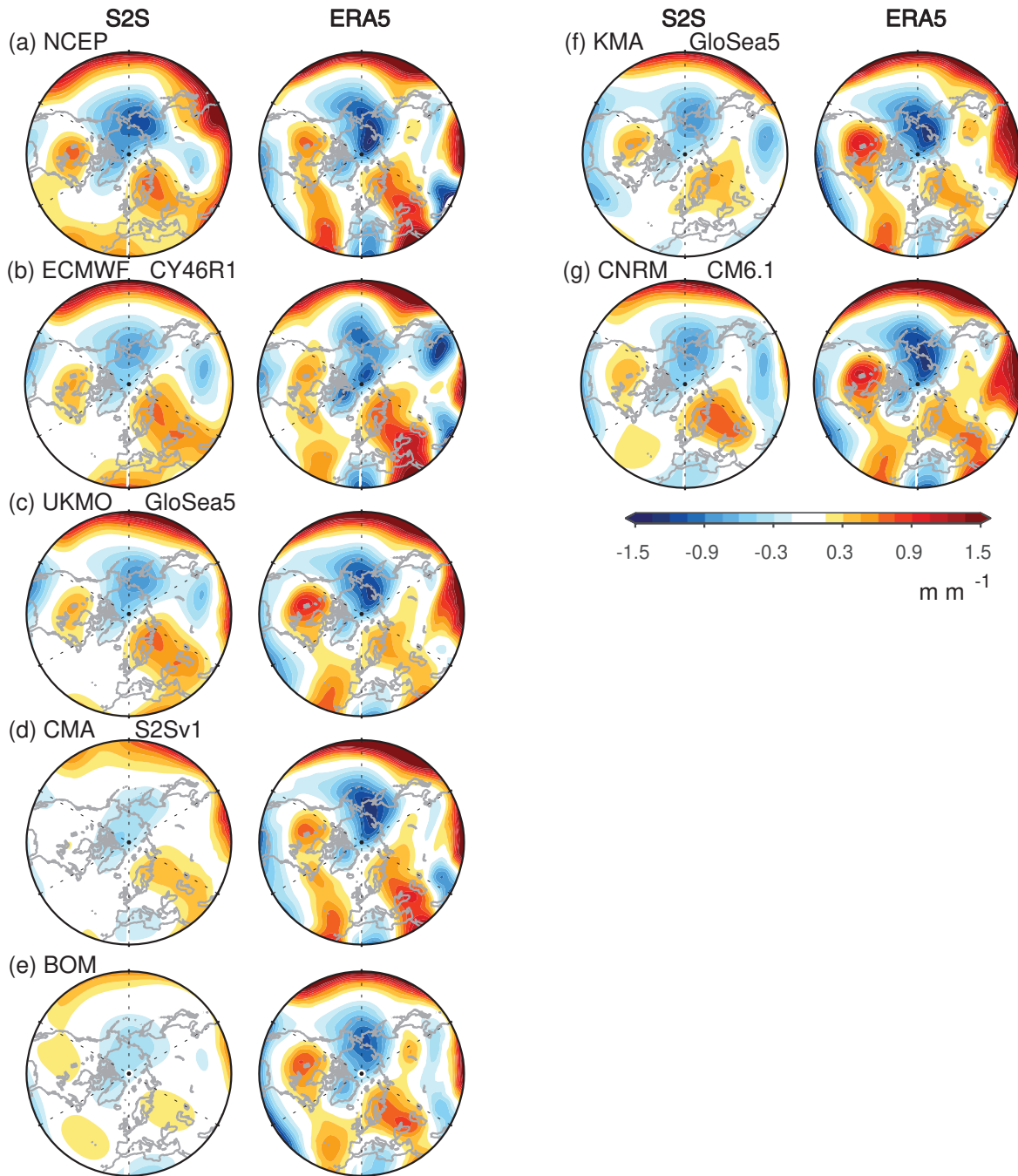
Figure 9. Regression coefficient of the change in polar cap geopotential height at 10-hPa over 10 days (Zpole,day10-Zpole,day0; a vortex weakening index) with Z500 anomalies on day 0. Uses week 3 and 4 of December/January initializations for Z500 anomalies on day 0.

The sensitivity of the SH polar vortex to 100-hPa extratropical heat flux is overestimated in most models, however the absolute error is higher for low-top models (Figures 4b and 8b). The ability of the vortex to modulate its own wave-driving is less pronounced in the SH than in the NH, however the models underestimate this effect by 28% (Figure 8).

3.3 Biases in interannual variance of daily heat flux extremes

Sections 3.1 and 3.2 demonstrated that there are systematic biases in heat flux variance and the associated upward coupling at subseasonal timescales. This bias also extends to a poor simulation of interannual variability of daily heat flux extremes. We quantify this behavior by computing the 95th percentile of daily eddy heat fluxes (wavenumbers 1-3) for each winter of the 1999-2010 period (Figures 10 - 11). The median (marker) and two standard deviation range (whiskers) of those values for each lead time are shown at 50 and 300 hPa. This analysis thus shows year-to-year spread in the highest heat flux extremes. An equivalent analysis was done for the 5th percentile (lowest) extremes with qualitatively similar results (not shown).

For the NH (Figure 10), the interannual spread in positive heat flux extremes at 50 hPa becomes dramatically reduced for most systems after week 1 compared to reanalysis. In other words, the year-to-year variations in stratospheric heat flux extremes are not well captured in the S2S forecast systems beyond a week. This contrasts with the behaviour at 300 hPa, where most forecast systems capture the reanalysis interannual spread in extremes through week 4 (days 22-28), though there is some reduction in spread by week 5. BoM has persistently too low positive heat flux extremes at both 50 and 300 hPa.

For the SH (Figure 11), systems underestimate the interannual spread of daily heat flux extremes beyond week 1 at 50 hPa and beyond week 2 at 300 hPa. This reduction in the spread of the positive heat flux extremes is particularly evident at 50 hPa, despite most systems capturing the median values of the 95% percentile extremes well (except for BoM which underestimates the median extreme value after week 1, and WACCM which underestimates the median after week 4). At 300 hPa, most systems show a reduction in interannual spread of positive heat flux extremes at week 3-5 compared to the reanalysis spread (and systematically underestimate the median extreme value). This attenuation in the spread of extreme values is thus more evident in the SH troposphere compared to the NH troposphere.

These analyses suggest that at long lead times, the models' daily heat flux extremes are either less sensitive to or lack external sources of interannual variability that arise due to, e.g., teleconnections, or are missing certain internal processes that lead to variability on longer timescales. While this reduction in spread is also apparent for median values of heat fluxes in some models (not shown), it is much weaker, suggesting that the extremes of the eddy heat flux distribution are more sensitive to this bias than the median.

3.4 Downward propagation of stratospheric polar vortex anomalies within the stratosphere

The downward propagation of mid-stratospheric polar vortex anomalies to the lower stratosphere is considered in Figure 12, which shows the lagged correlation of 10-hPa polar cap height with 100-hPa polar cap height. In the NH, downward propagation peaks after 2 days in ERA5 and in most models. While several models simulate this downward propagation realistically, there is a systematic underestimation of the magnitude of downward coupling within the stratosphere (Figure 4a, fourth row), with only 2 of 21 systems (low-top BoM and CMA-S2Sv1) simulating a too-strong coupling strength. Biases are even more pronounced
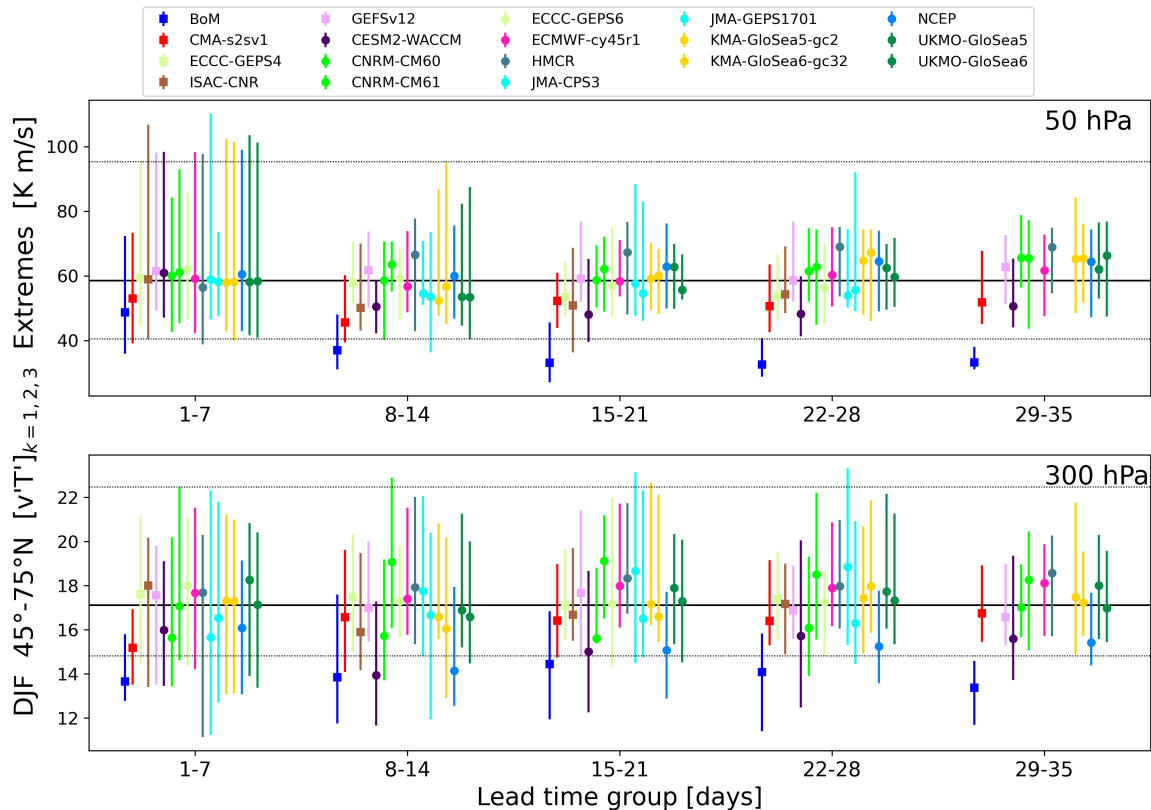
Figure 10. The 95th percentile daily eddy heat flux extremes for 45-75°N during winter (DJF) from 1999-2010 for all models by weekly lead time group. The median is indicated by the marker and the ±2 standard deviations by the whiskers for (top) 50hPa and (bottom) 300hPa. The equivalent values from reanalysis are given by the horizontal black lines (bold: median; thin: ±2 standard deviations).

for lags of 5 to 10 days, though smaller for 20 day lags aside from low-top models (Figure 12). BoM again shows the largest bias (specifically, an overestimation of coupling strength), even if its correlation indicates an underestimation of coupling strength; this is again a reflection of a poor simulation of stratospheric variance (Figure 2a). There is a notable improvement from CNRM-CM60 to CNRM-CM61, from UKMO-GloSea5 to UKMO-GloSea6, and in successive versions of ECCC-GEPS, though not from ECMWF-cy45r1 to ECMWF-cy47r3 or JMA-GEPS1701 to JMA-CP3 (Figure 4a). Low-top models (Figure 5d) and models with relatively poor climatological stationary waves tend to simulate a stronger downward coupling strength; however, this relationship is dominated by a single outlier model (BoM). If the correlation is computed without this model, there is instead no detectable relationship between downward coupling strength and either model lid height or stationary wave amplitude.

Similar to the NH, downward coupling of polar cap height from the mid- to lower-stratosphere is too weak in the SH in nearly all models (Figure 12b, 4b), especially at longer lags. Notably, in the SH the reanalysis relationship actually strengthens
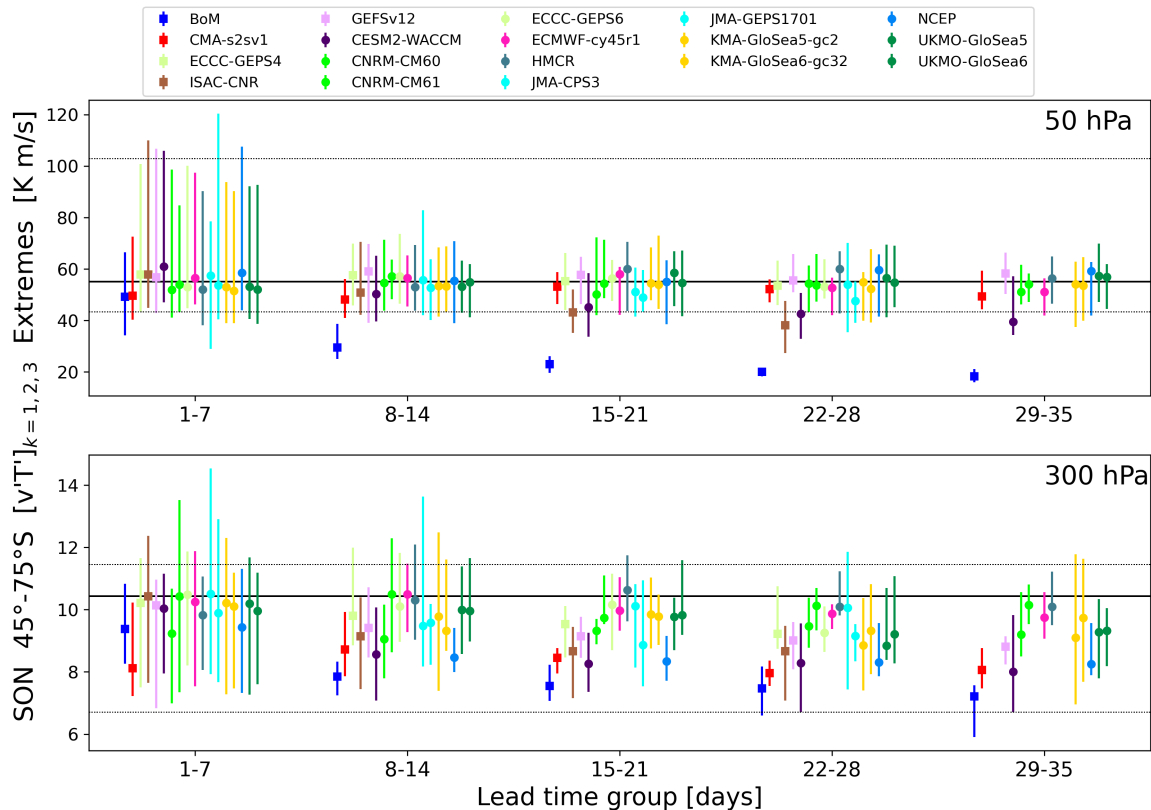
Figure 11. The 95th percentile daily eddy heat flux extremes for 45-75°S during spring (SON) from 1999-2010 for all models by weekly lead time group. The median is indicated by the marker and the ±2 standard deviations by the whiskers for (top) 50hPa and (bottom) 300hPa. The equivalent values from reanalysis are given by the horizontal black lines (bold: median; thin: ±2 standard deviations).

between days 4-20, which may be related to chemistry-circulation coupling in austral spring, as discussed by Simpson et al. (2011). High-top models overall perform better and have a lower absolute error.

3.5 Persistence of the polar vortex signal in the lower stratosphere and downward propagation from the lower stratosphere to the near-surface

After the stratospheric signal reaches the lower stratosphere, it can subsequently impact the tropospheric circulation. We evaluate whether the models successfully capture this effect using both a regression/correlation approach and a compositing approach.
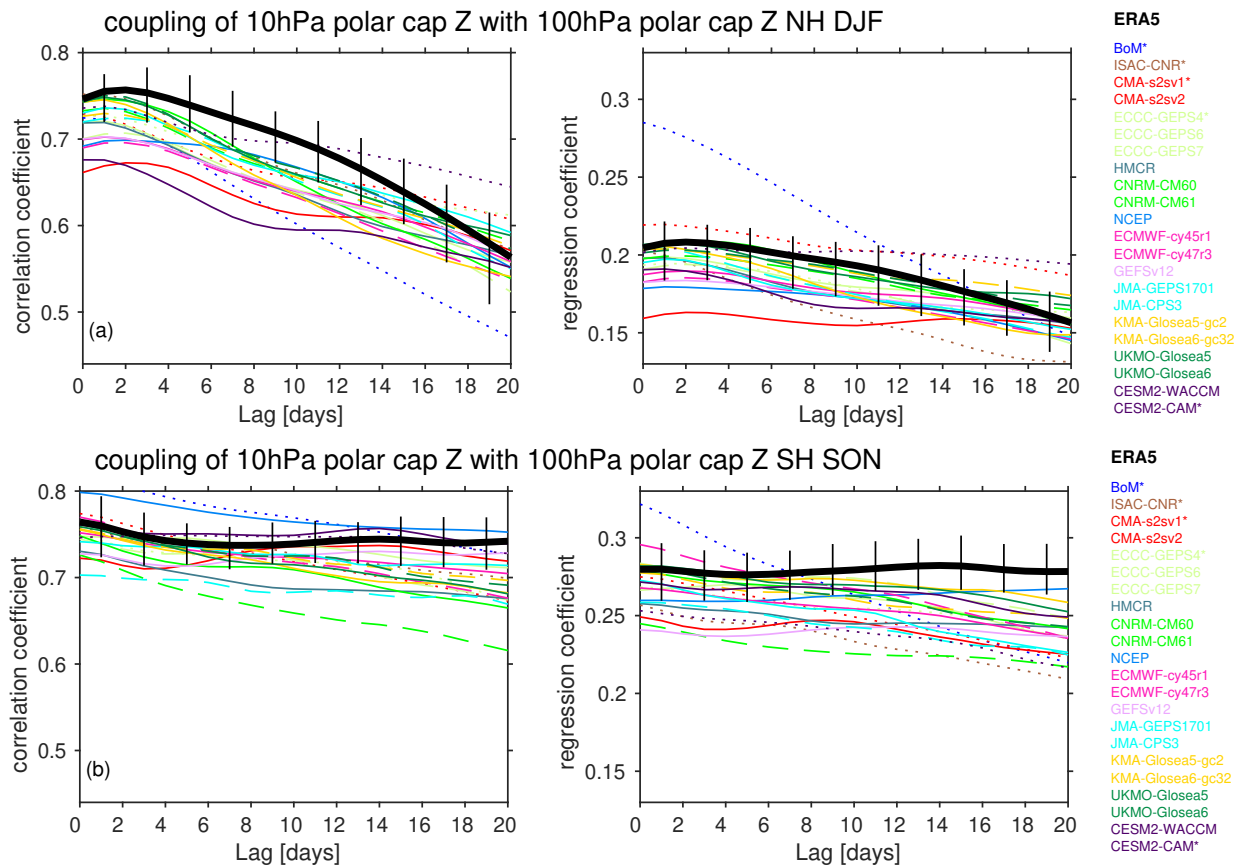
Figure 12. Coupling of polar cap height at 10-hPa with that at 100-hPa. Polar cap Z at 10-hPa is taken from days 9 to 12, and we range 100-hPa polar cap Z from simultaneous with Z10 to 20 days after. Low top models are dotted. Older versions of high-top models are dashed. Black vertical lines show the range in coupling strength upon subsampling ERA5 reanalysis to match each of the forecast systems, and the solid black line indicates the mean of these coupling strengths.

3.5.1 Regression/correlation perspective on downward coupling biases

We begin with a regression/correlation approach in Figure 13a, which shows the lagged regression of 100-hPa polar cap height with 850-hPa polar cap height in the NH. For lags of less than a week, too-strong biases exceeding 5% are evident in 12 models, while too-weak biases exceeding 5% are evident in only three models (ISAC-CNR, JMA-GEPS1701, and NCEP). The too-strong downward coupling for near-simultaneous lags is consistent with Kolstad et al. (2020) for ECMWF. For later lags, additional models develop too-weak biases, and individual models suffer from large biases. For example, CESM2-WACCM, CESM2-CAM, and CNRM (both generations) overestimate the coupling, while JMA (both generations) and NCEP underestimate it. There is a substantial improvement from ECMWF-cy45r1 to ECMWF-cy47r3 and from ECCC-GEPS6 to ECCC-GEPS7, but we see no evidence for an improvement from the other modeling centers.
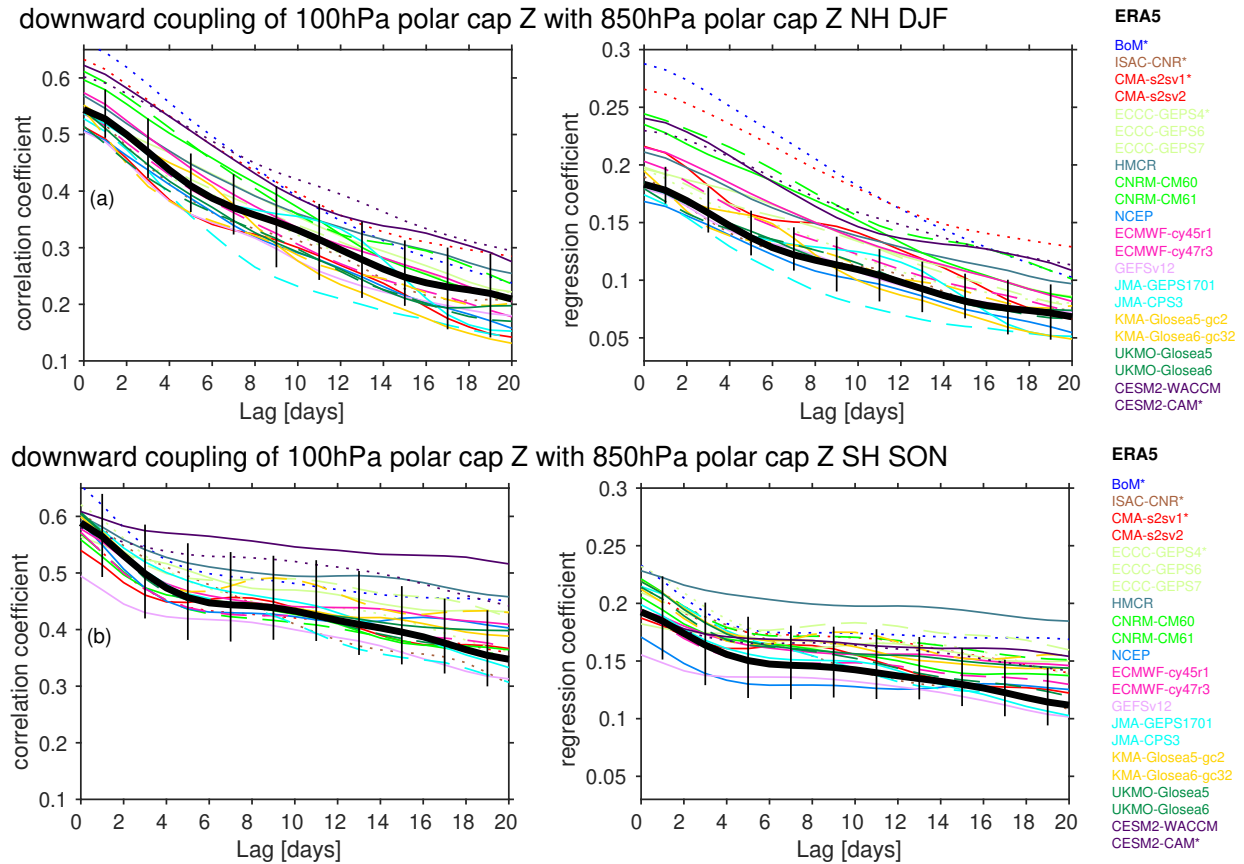
Figure 13. Lagged correlation/regression coefficient of 100hPa polar cap geopotential height with that at 850hPa. Polar cap geopotential height at 100hPa is selected from days 9 to 12 after initialization, and we range 850hPa polar cap Z from simultaneous with Z100 to 20 days after. Low top models are dotted. Older versions of high-top models are dashed. Black vertical lines show the range in coupling strength upon subsampling ERA5 reanalysis to match each of the forecast systems, and the solid black line indicates the mean of these coupling strengths.

Compared to other stratosphere-troposphere coupling metrics (Figure 4a), this part of the coupling process is the most
315 consistently biased (in an absolute sense) across models. The bias is less evident upon examining the correlation (Figure 13, Supplemental Figure S2), likely because some of these models also suffer from too-strong biases in the variance for 850-hPa geopotential height (Figure 2a): a too-strong regression coefficient combined with too-strong variance can lead to a reasonable correlation in the net. These smaller biases for a correlation approach are consistent with Lee and Charlton-Perez (2024) for models which overlap with those considered here (ECMWF-cy45r1, CNRM-CM60, UKMO-Glosea5, and NCEP). Downward
320 coupling is too strong in models with overly weak climatological tropospheric wave-1 (Figure 6f); this relationship is consistent with the documented effect of planetary waves to dampen synoptic eddy feedback (i.e. Feldstein and Lee (1998); Lorenz and Hartmann (2003), though the full range of interactions of planetary waves with vortex perturbations is still not fully understood (Song and Robinson, 2004; Domeisen et al., 2013; Hitchcock and Simpson, 2014; Garfinkel et al., 2023).
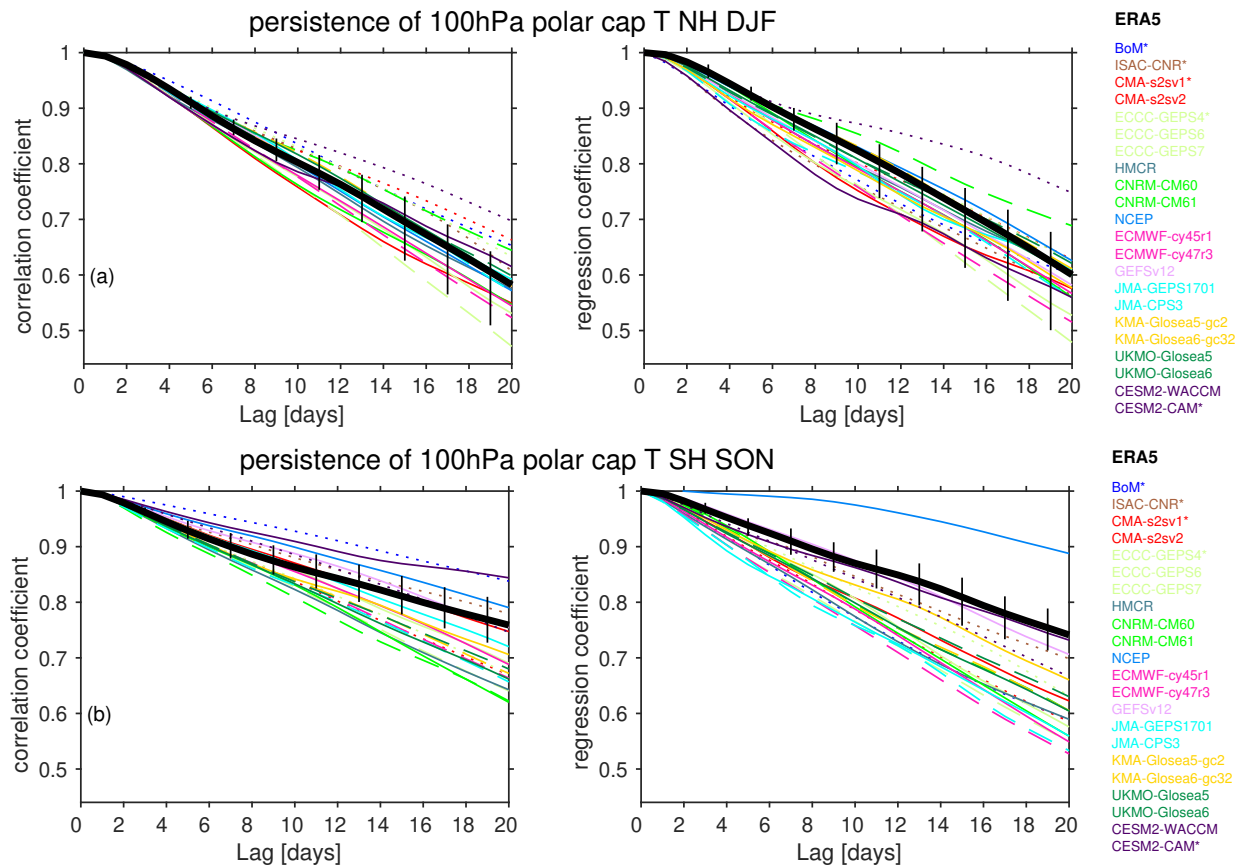
Figure 14. Persistence of 100hPa polar cap T. Polar cap T at 100hPa is taken from days 9 to 12, and we then compute its lagged correlation up to 20 days later. Low top models are dotted. Older versions of high-top models are dashed. Black vertical lines shows the range in coupling strength upon subsampling ERA5 reanalysis to match each of the forecast systems, and the solid black line indicates the mean of these coupling strengths.

The downward coupling signal in later weeks is potentially related to the persistence of lower stratospheric vortex anomalies, as the slow radiative decay of these anomalies allows for lower stratospheric variability to affect surface climate on subseasonal timescales (Hitchcock and Simpson, 2014). Specifically, if polar vortex anomalies were to decay too fast, then this could lead to a too-weak downward coupling at later lags. This possibility is examined in Figure 14, which shows the lagged auto-correlation of polar cap temperature at 100-hPa; we focus here on temperature rather than geopotential height due to its close connection with radiative timescales and tracer concentrations. Three models simulate biases of the autocorrelation of polar cap temperature at 100-hPa on day-20 exceeding 5% (low-top CESM2-CAM and BoM, and high-top NCEP). Seventeen other models simulate overly-fast decay if we subsample ERA5 to match the dates actually used for each model (Figure 4). The overly-fast decay exceeds 10% only for ECCC-GEPS6, ECCC-GEPS7, CNRM-CM6.1, and JMA-GEPS1701, and is more pronounced (though not statistically significantly so) in models with higher tops and better stationary waves (Figure 5e and

6e). The correlation between the auto-correlation of polar cap temperature at 100-hPa on day-20 with the regression of 100-hPa polar cap height with 850-hPa polar cap height on day-20 is 0.34, such that a stronger autocorrelation of polar cap temperature is associated with a strong surface signal. This relationship is somewhat weaker than the corresponding relationship with tropospheric stationary waves (r=-0.45).

In the SH, downward coupling of polar cap height from the lower stratosphere to the surface is too strong in most models for both near-instantaneous and at 20-day lags (Figures 13b, 4b). Two models (NCEP and GEFSv12) show too weak coupling for simultaneous lags exceeding 10%, and at later lags ISAC-CNR and JMA-CPS3 also simulate too-weak coupling. For nearly all other models, however, overly strong downward coupling occurs even as polar cap temperature anomalies decay too fast in these models (Figures 4b, 14b). Hence the too-strong downward coupling likely reflects overly strong eddy feedback, as has been recently shown explicitly for a subset of these models (Garfinkel et al., 2024). Consistent with this, the too strong coupling bias is more pronounced at later lags than near-simultaneous lags (Figure 4b).

3.5.2 Extreme stratospheric events perspective on downward coupling biases

So far our consideration of downward coupling has been based on a correlation/regression analysis. This analysis does not explicitly consider the role of extreme events of the stratospheric polar vortex for surface predictability. Specifically, a highly disturbed or extremely strong polar vortex has stronger impacts than more typical vortex variability; for example, White et al. (2020), White et al. (2022), and Garfinkel et al. (2023) find the near-surface response to scale linearly with the lower stratospheric perturbation. We now consider whether the S2S systems capture downward coupling for these extreme events.

We quantify the biases in downward impact by forming composites of initializations in which polar cap height anomalies at 10hPa exceed 500m (strong vortex) or are more negative than -500m (weak vortex) on day 10, and compute the Zcap at 100hPa on days 10 through 31 (Supplemental Figure S5). These thresholds are chosen to consider extreme conditions only (approximately 9% of all available members are included in each composite), though results are similar for a threshold of, say, ±400m (not shown). The biases averaged from days 20 to 30 are summarized in Figure 15a. For both the SH and NH, many more models simulate too-weak downward propagation within the stratosphere than too-strong. This effect is consistent with the regression coefficients (Figure 4 and 12). The bias is particularly pervasive for weak vortex events.

Next, we consider biases in the downward coupling of extreme vortex events between 100hPa and the near-surface. Specifically, we form composites of initializations in which polar cap height anomalies at 100hPa exceed 175m (strong vortex) or are more negative than -175m (weak vortex) on day 10, and plot the Zcap at 850hPa on days 10 through 31 (Supplemental Figure S6). These thresholds are chosen such that $\sim 7.6\%$ of all available members are included in each composite, though results are similar for a threshold of, say, ±100m (not shown). The biases averaged from days 20 to 30 are summarized in Figure 15b. In contrast to the too-weak downward propagation within the stratosphere, most models simulate too-strong downward coupling from the lower stratosphere to the near-surface. There are notable exceptions, however, in both hemispheres. In the NH, ISAC-CNR, both JMA, and all four GloSea configurations are relatively less biased, consistent with the too-weak regression coefficients evident in Figure 4. Similarly, in the SH, ISAC-CNR, NCEP, GESFv12, and both JMA simulate too weak coupling, also consistent with Figure 4. Overall, downward coupling is too-weak within the stratosphere, but too-strong from the lower
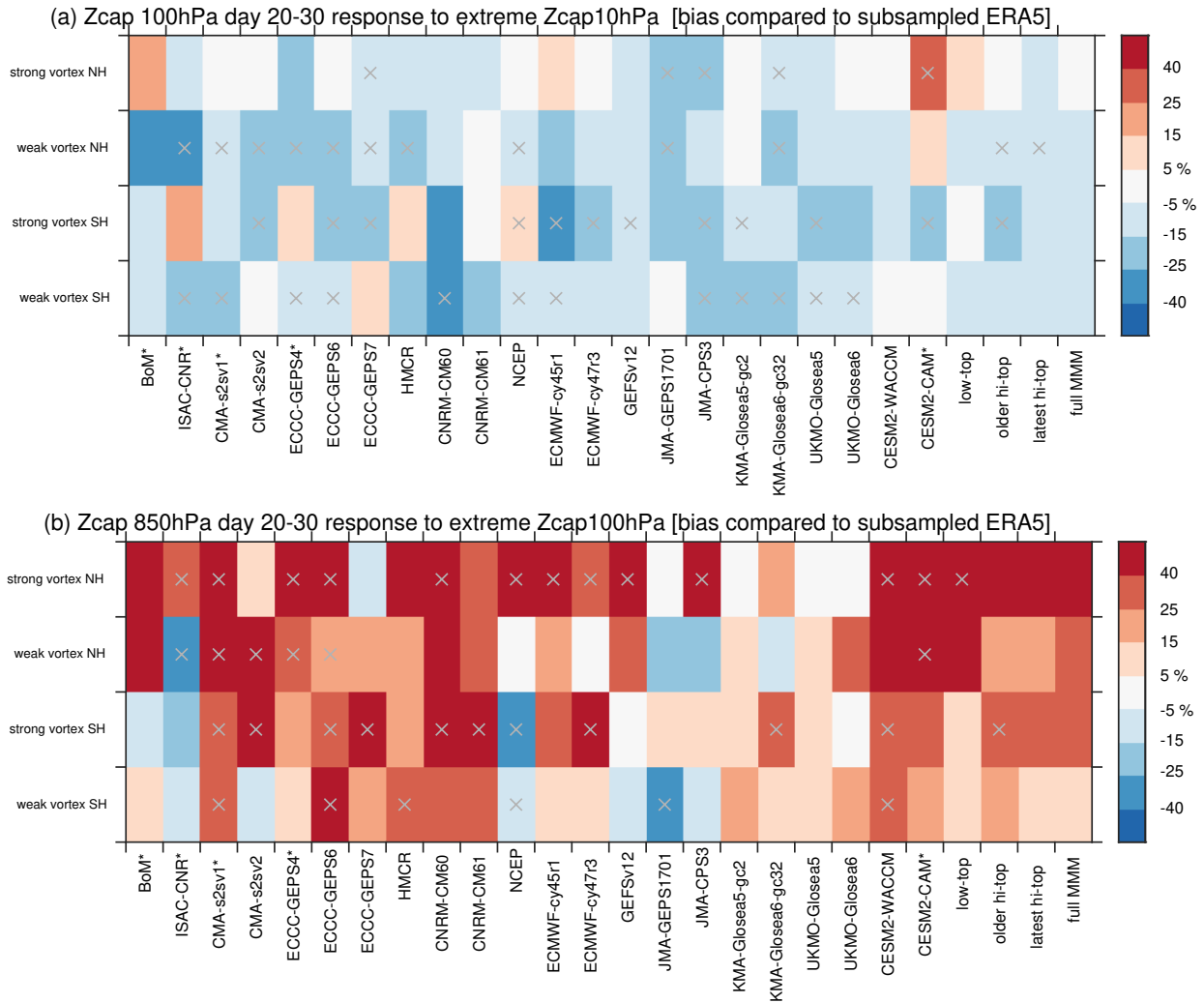
Figure 15. Summary of the biases in downward coupling strength for extreme stratospheric events. For each forecast system we compare to the corresponding identical period in ERA5, and then show the percentage error. The bias is defined as the difference between the model and ERA5 divided by the response in ERA5. A gray '×' indicates models and metrics for which all ensemble members simulate a bias in the coupling strength of the same sign, or alternatively if ERA5 doesn't fall within the envelope of the available members. Low-top models are denoted with an asterisk after their name. (a) downward coupling within the stratosphere: (first, third row) composite mean of Zcap at 100hPa for days 20-30 for initializations in which Zcap at 10hPa on day 10 exceeds 500m; (second, fourth row) as in first and third rows but Zcap at 10hPa on day 10 is more negative than -500m. (b) downward coupling from the lower stratosphere to the troposphere: (first, third row) composite mean of Zcap at 850hPa for days 20-30 and initializations in which Zcap at 100hPa on day 10 exceeds 175m; (second, fourth row) as in first row but Zcap at 100hPa on day 10 is more negative than -175m. The thresholds lead to $\sim 8.7\%$ of all available members being chosen averaged across all models, hemispheres, and composites; the mean composite size is 248.

stratosphere to the near-surface. These two biases tend to compensate each other when considering downward coupling from the mid-stratosphere to the near-surface. Indeed, a similar figure but for the Zcap 850hPa response to extreme events of Zcap at 10hPa shows lower biases (less than 3% and 9% in the SH and NH, respectively) than for Zcap at 100hPa.

3.6 Other possible contributors to inter-model spread in coupling strength

Throughout the text we discussed the role of model lid height and of biases in the stationary waves for biases in coupling processes. These S2S models are also known to suffer from a cold pole bias in the lowermost stratosphere (Lawrence et al., 2022), and we have explored whether intermodel spread in the magnitude of this cold pole bias might be related to spread in the strength of coupling. Supplemental figure S7 and S8 consider this possibility; however, we find that its role is weaker than those of lid height or stationary waves for all coupling processes.

An additional possibility is that the number of levels in the troposphere, near the tropopause, or in the stratosphere, might be related to the model spread in coupling processes. We evaluate these possibilities with figures S9, S10, and S11 respectively. Namely, we contrast coupling strength with the number of model levels below 300hPa, between 300hPa and 100hPa, and between 100hPa and 10hPa. The number of tropospheric levels is robustly associated with improved coupling of wave-1 heat flux from 500hPa to 100hPa (Figure S9a), and the correlation (0.67) is stronger than for any of the other factors explored in this paper (e.g., model lid, stationary wave climatology, cold pole biases). A similar albeit weaker relationship is evident if we consider the number of levels from 300hPa to 100hPa (Figure S10a, r=0.44). This sensitivity is not evident in the SH or for wave-2, however. More levels in the troposphere and near the tropopause are associated with a more reasonable sensitivity of the polar vortex to incoming wave flux in the NH (Figure S9c, S10c), however this effect is weaker than the corresponding effect if we consider the climatological heat flux bias (Figure 6c). There is little relationship between downward propagation within the stratosphere and the number of levels if BoM is excluded (Figure S9dj, S10dj, S11dj), however downward propagation from the lowermost stratosphere to the near-surface in the NH is better simulated in models with more tropospheric levels (Figure S9f). Finally, the number of levels between 100hPa and 300hPa (i.e., better resolution in the tropopause region) is robustly related to better relaxation timescales for Tcap100 in the Northern Hemisphere (Figure S10e). Such a relationship makes sense if these relaxation timescales are dictated by poor representation of transport of water vapor (Riese et al., 2012; Charlesworth et al., 2023).

For essentially all of the coupling processes the number of model levels between 100hPa and 10hpa is not significantly correlated to biases in the coupling processes (Figure S11). This suggests that better simulating/resolving tropospheric and lower stratospheric processes is the key to reducing some of the biases in the coupling that we are seeing, rather than getting extra-high resolution higher in the stratosphere. We have tested this possibility by re-calculating the metrics in Figure 5 but for high top models only, and found that there is still a tendency for higher high-top models to better represent upward wave-2 (not shown). However, for other metrics the correlations are reduced or even change sign, suggesting that once lid is sufficiently high, the effect of lid height becomes saturated.

4   Discussion and Conclusions

Variability in the extratropical stratosphere and troposphere are coupled (Baldwin and Thompson, 2009; Kidston et al., 2015). A large pulse of planetary wave in the troposphere can disturb the polar stratospheric vortex, while vortex extremes influence surface climate and extremes for weeks to months afterwards (Domeisen and Butler, 2020). This coupling can potentially provide windows of opportunity for prediction on subseasonal-to-seasonal (S2S) timescales (Butler et al., 2019; Domeisen et al., 2020b), however model biases in either the troposphere or stratosphere can degrade the representation of these coupling processes.

We have performed a comprehensive intercomparison of biases in extratropical stratosphere-troposphere coupling processes in subseasonal forecast systems, with a core focus on systems that contribute to the S2S database (Vitart et al., 2017). We broke up this coupling into six processes that can be diagnosed with a few key metrics, in the hopes that they can be easily adopted by modellers to assist ongoing development. Our main results can be summarized as follows:

1. **Upward flux of wave activity to the lower stratosphere**

   NH: The forecast systems systematically underestimate the upward coupling of wave-1 from the mid-troposphere to the lower stratosphere. In contrast, upward coupling of wave-2 is better simulated (Figure 3; top two rows of Figure 4a). Upward coupling is better captured in high-top models, and even more robustly, in models with a better representation of climatological quasi-stationary waves (Figure 5a, 6a) and better tropospheric resolution (Figure S9). Models underestimate the sensitivity of lower stratospheric wave-1 heat flux to tropospheric variability in the Northwest Pacific and Euro-Atlantic (Figure 7). This relatively better performance for wave-2 as compared to wave-1 is remarkable given the overall poorer performance of these models with respect to the prediction of SSW events dominantly driven by wave-2 (Taguchi, 2018; Domeisen et al., 2020b). This difference between wave-1 and wave-2 biases in the upward wave flux is likely a reflection of the fact that climatological wave-2 heat flux is better represented (and indeed, too strong) in many of these models while climatological wave-1 is too-weak (Supplemental Figure S12). However, it is possible that there are additional biases in wave-2 ahead of extreme vortex events.

   SH: The high-top forecast systems systematically overestimate the upward coupling of wave-1 from the mid-troposphere to the lower stratosphere (Figure 3; top row of Figure 4b), in contrast to the underestimation in the NH. Note that the models also better capture the climatological wave-1 in the SH than in the NH (Supplemental Figure S12), and the intermodel spread in upward coupling in the SH is also linked to each model's representation of climatological wave-1 (Figure 6g). The tropospheric heat flux variances are systematically too high while the stratospheric variances are too low, so the relatively successful coupling strength may be due to some kind of cancellation effect (too high variability in the troposphere is overcompensating for what would be a too-weak upward coupling).

2. **Sensitivity of the vortex to upward flux of wave activity in the lower stratosphere**

   NH: The polar vortex is not sensitive enough to upward propagating wave flux (Figure 8a). This effect is especially pronounced in models with large biases in climatological 500-hPa heat flux (Figure 6c).

SH: Multi-model mean biases are small (Figure 8b). The intermodel spread is mostly accounted for by the climatological 500-hPa heat flux (Figure 6i). Note that the forecast systems simulate climatological 500-hPa heat flux better in the SH than in the NH in the multi-model mean.

3. **Interannual variance of daily heat flux extremes**

In both the NH and SH stratosphere, the interannual spread in positive eddy heat flux extremes is strongly reduced for most systems after week 1. This is also evident in the SH troposphere for weeks 3-5. More work is needed to understand what drives this lack of interannual variability in heat flux extremes (one possibility might be poor simulation of teleconnections arising from e.g., ENSO; Garfinkel et al., 2022; Bayr et al., 2019; Williams et al., 2023), as well as the asymmetry in behavior between the NH and SH troposphere, and the extent to which this bias affects stratospheric circulation extremes, their predictability, and subsequent downward coupling. Potential implications for subseasonal forecasting include, for example, a failure of the S2S systems to forecast stratospheric heat flux extremes beyond week 1 that are associated with potentially predictable sources of interannual variability.

4. **Downward propagation within the stratosphere**

NH: There is a systematic underestimation of the magnitude of downward coupling within the stratosphere both when using a regression/correlation approach (Figure 4a, 12a) or a compositing approach focused on the extreme events (Figure 15). We were unable to identify any factor that is robustly linked to the intermodel spread in this underestimation (Figure 5d, 6d, S7d-S11d).

SH: Similar to the NH, downward coupling of polar cap height from the mid- to lower-stratosphere is too weak in the SH in nearly all models (Figure 4b, 12b), especially at longer lags, however the biases are generally small (<10%). This finding is confirmed using a composite approach based on extreme events (Figure 15). As for the NH, we were unable to identify any factor that is robustly linked to the intermodel spread in this underestimation (Figure 5j, 6j, S7j-S11j).

5. **Persistence of the polar vortex signal in the lower stratosphere**

NH: The multi-model mean bias for high top models is less than 5%, however there is a wide spread across models with too-strong persistence for some models and too-fast decay (albeit relatively weak) for most models. We have examined whether intermodel spread in this bias is related to mean state biases in polar cap temperatures, however the relationship was weak (Supplemental Figure S7, S8). The intermodel spread in this underestimation isn't related to model lid or stationary wave climatology either (Figure 5e, 6e), however it is related to the number of vertical levels between 100hPa and 300hPa (Figure S10e).

SH: Temperature anomalies decay too fast. Models with more levels between 100hPa and 300hPa tend to suffer from this problem more severely (Supplemental Figure S10k), suggesting that adding resolution is not a panacea. The intermodel spread in this bias is related to mean state biases in polar cap temperatures: models with larger mean-state cold biases simulate a better auto-regression (Supplemental Figure S7, S8). Possible speculative causes for this include (i) a stronger time-mean vortex can better duct away incoming waves, and hence is less variable; (ii) a cold bias will lead to less

efficient longwave cooling in response to a temperature anomaly (regardless of sign); (iii) a third, as yet unknown, bias may also be important. An additional possibility is that ozone coupling may be crucial for temperature persistence in the SH, however ozone is prescribed to climatological values in many subseasonal forecasting models. It is notable that NCEP is the only model overpredicting persistence in the SH and that is one of the few models used in this study that uses prognostic ozone. Additional work is needed to better understand this possibility.

6. **Downward propagation from the lower stratosphere to the near-surface**

NH: Downward coupling is too strong at both short and longer lags (Figure 4a, 13, 15), for both a regression approach and a composite approach based on extreme events. In contrast, a correlation approach indicates that biases are relatively small in the multi-model mean (Figure 13, Supplemental Figure S2, consistent with Lee and Charlton-Perez (2024)). This difference in the overall conclusion as to whether downward coupling is biased among the different methodologies is likely due to too-strong variance in Zcap850 in most models (Figure 2). Regardless of methodology, downward coupling from the mid-stratosphere to the near-surface is of reasonable strength in the multi-model mean. The multi-model mean coupling strength is the net effect of qualitatively different behaviors across models, however, and this metric is the most biased (in an absolute sense) of any across models. Downward coupling is stronger in models with poor climatological stationary waves, low tropospheric vertical resolution, or with too long a persistence timescale of lower stratospheric temperature anomalies (Figure 6f and S9f). This sensitivity to climatological stationary waves is consistent with the known damping on annular mode variations provided by planetary waves (Feldstein and Lee, 1998; Lorenz and Hartmann, 2003), though planetary waves may couple with vortex perturbations directly and act to bring vortex perturbations down to the surface (Song and Robinson, 2004; Simpson et al., 2013; White et al., 2020).

SH: Downward coupling of polar cap height from the lower stratosphere to the surface is too strong in most models (Figures 4b, 13b, 15) even as polar cap temperature anomalies decay too fast in these models (Figures 4b, 14b). Hence the too-strong downward coupling likely reflects overly strong eddy feedback in the SH (while the NH eddy feedback has an opposite signed bias, namely it is too-weak), as has been recently shown for a subset of these models (Garfinkel et al., 2024).

The results above are based on relatively short hindcast periods, so that the ERA5 correlations/regressions shown throughout may be subject to sampling variability. Indeed, Lawrence et al. (2023) showed that similar coupling metrics in GEFSv12 largely fell within the range of ERA5 sampling variability. Here we assume that since the S2S systems are initialized with the same internal variability as observed in the real world, and are intended to be useful for predicting on subseasonal timescales, that they should be able to reproduce the ERA5 values (subsampled for each hindcast), and documenting the deviations from these values particularly in a multi-model comparison still enhances understanding of where and how the models are biased. Nonetheless, some of the model biases shown here fall within the range of ERA5 sampling variability (which can be estimated using the black vertical bars on, e.g., Figures 12 and 13).

The NH polar vortex in these forecasting models is insufficiently coupled to tropospheric variability, consistent with the too-weak impact of predictable tropospheric modes of variability such as the Madden Julian Oscillation and snow cover anomalies

on the vortex documented in previous work using a subset of these models (Domeisen et al., 2020b; Garfinkel et al., 2020; Schwartz and Garfinkel, 2020; Stan et al., 2022). This conclusion is consistent with Lee et al. (2020), who also found that models systematically underestimate the stratospheric heat flux and vortex response to a Ural blocking-like pattern. In contrast, the SH stratospheric vortex is coupled realistically with tropospheric variability. Interestingly, older generation of chemistry-climate models analyzed by Eyring et al. (2006) displayed the correct stratospheric response of polar temperatures to wave forcing in the Northern, but not in the Southern Hemisphere. However, their conclusions are based on 20 years of seasonal mean data in free-running atmospheric simulations without an attempt to rigorously quantify uncertainties. Here, we are focusing on shorter timescales, initialized forecasts, and have orders of magnitude more data per model, which allow for a more stringent criteria of fidelity.

Downward coupling from 100hPa to 850hPa is too strong in both hemispheres in the multi-model mean, though a few models have an opposite signed bias (e.g., NCEP, JMA-CP3, and ISAC-CNR). While we link this in our study to biases in synoptic eddy feedback, persistence of lower stratospheric temperature anomalies, and quasi-stationary waves, there are other possible causes that might be relevant. Specifically, stratospheric ozone-circulation coupling is crucial in the SH spring and summer, and also has an important role in the NH spring. Some studies have shown that using prescribed ozone that includes year-to-year variations instead of climatological ozone improves SH forecast skill of surface climate (Hendon et al., 2020; Oh et al., 2022). Experiments with fully interactive ozone show further improvements in the representation of the tropospheric response (Romanowsky et al., 2019; Friedel et al., 2022a, b), although the downward coupling in models with interactive ozone is also strongly affected by model biases (Bergner et al., 2022). Future work should explore the role of prognostic or interactive ozone in S2S operational systems for downward coupling and improvements in predictive skill.

A poor representation of gravity waves is known to degrade stratosphere-troposphere coupling (Shaw and Perlwitz, 2010; Wicker et al., 2023), however the S2S archive does not include gravity wave drag as a standard output and even models with ostensibly similar parameterizations can nonetheless have very different net fluxes (Lott et al., 2024). Future work should evaluate the role of gravity waves for coupling strength should the requisite output be made available.

We find that the models better capture wave-2 vertical coupling from 500hPa to 100hPa, likely because the biases in their climatological wave-2 heat flux are smaller than for wave-1. This appears to be contrary to climate models, which struggle more with wave-2 in the NH (it is typically too weak) and also tend to overestimate the number of wave-1 SSWs with respect to wave-2 events. Nevertheless, there is a notable bias in coupling of wave-2 between 100hPa and 500hPa at negative lags in the NH (second row of Figure 3, lags -6 to 0). Specifically, strong values of 100hPa heat flux have a weak tendency to precede pulses at 500hPa, however only one model captures this effect. This may reflect problems more generally with downward wave coupling and/or wave reflection; exploring this possibility in greater detail is left for future work.

We have formulated a reduced set of key metrics and diagnostics that can be saved and analyzed relatively easily as part of the model development cycle. We hope this set of diagnostics will be adopted and will aid the development of improved models. We also want to emphasize that this analysis is only possible with the output of stratospheric data. The relative paucity of levels makes it difficult to more fully diagnose why the upward coupling strength and downward coupling strength within the stratosphere is too weak in most models. For example, this bias could be related to biases in the representation of the

535   tropopause and lowermost stratosphere (Weinberger et al., 2022), however such an effect is impossible to diagnose with data only at 200hPa, 100hPa, and 50hPa. Finally, the implications of poor coupling for surface climate and predictability in specific regions where the stratosphere is known to have a large impact need to be explored.

Data availability. The hindcasts from the S2S database used here are available from https://apps.ecmwf.int/datasets/data/s2s/ under the "Reforecasts" S2S set. The NOAA GEFSv12 hindcasts can be obtained from https://registry.opendata.aws/noaa-gefs-reforecast/. Hindcasts
540   for CESM2-CAM are available at https://www.earthsystemgrid.org/dataset/ucar.cgd.cesm2.s2s_hindcasts.html, while those for CESM2-WACCM are from https://www.earthsystemgrid.org/dataset/ucar.cgd.cesm2-waccm.s2s_hindcasts.html.

Author contributions. CIG and AHB drafted the paper. CIG produced the final version of all figures except Figure 10 and 11. ZDL organized and led the SNAP effort leading to this paper, and also downloaded all of the data. AHB produced the final version of Figure 10 and 11. EDS produced an earlier version of Figure 10 and 11. IS and AYK produced an earlier version of Figure 13. GK produced earlier versions
545   of Figures 5 and 6. All the listed coauthors were active participants in this SNAP community effort and provided comments on the draft manuscript.

Acknowledgements. This work uses S2S Project data. S2S is a joint initiative of the World Weather Research Programme (WWRP) and the World Climate Research Programme (WCRP). This work was initiated by the Stratospheric Network for the Assessment of Predictability (SNAP), a joint activity of APARC (WCRP) and the S2S Project (WWRP/WCRP).
550   C.I.G and J.R. are supported by the ISF-NSFC joint research program (ISF grant No. 3065/23 and National Natural Science Foundation of China grant no. 42361144843). C.I.G. and J.C. are supported by the NSF-BSF joint research program (United States-Israel Binational Science Foundation grant no. 2021714 and National Science Foundation grant no. AGS-2140909). I.S. and A.Y.K are supported by the Research Council of Finland (grant No. 355792.) The work of M.O. is supported by UBACyT 20020220100075BA, PIP 11220200102038CO and
555   PICT-2021-GRF-TI-00498 projects. The work of A.dlC. is funded by the Spanish Ministry of Science through project PID2022-136316NB-I00. M.A., B.A and N.C. acknowledge the support of the Spanish Ministry of Science and Innovation through the RecO3very (PID2021-124772OB-I00) project. FMP and JG-S have been partially supported by the Spanish ATLANTE project (PID2019-110234RB-C21) and "Ramón y Cajal" programme (RYC-2016-21181), respectively. N.P.H. and C.J.W. are supported by UK Natural Environment Research Council (NERC) grant number NE/S00985X/1. C.J.W. is also supported by a Royal Society University Research Fellowship URF/R/221023.
560   S.-W.S. and H.K. are supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (2023R1A2C3005607). The work of R.W. is funded through ETH grant ETH-05 19-1. Support from the Swiss National Science Foundation through project PP00P2_198896 to D.D. is gratefully acknowledged.

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological & Environmental Research (BER), Regional and Global Model Analysis (RGMA) component of the Earth and Environmental System Modeling Program under Award Number DE-SC0022070 and National Science Foundation (NSF) IA 1947282. This work was also supported by the National Center for Atmospheric Research (NCAR), which is a major facility sponsored by the NSF under Cooperative Agreement No. 1852977. P.L. is supported by award NA18OAR4320123 from the National Oceanic and Atmospheric Administration (NOAA), U.S. Department of Commerce. ZDL was partially supported under NOAA Award NA20NWS4680051; ZDL and JP also acknowledge support from US Federally Appropriated Funds. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of NOAA, or the U.S. Department of Commerce.

References

Afargan-Gerstman, H., Jiménez-Esteve, B., and Domeisen, D. I. V.: On the relative importance of stratospheric and tropospheric drivers for the North Atlantic jet response to sudden stratospheric warming events, Journal of Climate, 35, 6453–6467, 2022.

Andrews, D. G., Leovy, C. B., and Holton, J. R.: Middle Atmosphere Dynamics, Academic Press, 1987.

575 Ayarzagüena, B., Palmeiro, F. M., Barriopedro, D., Calvo, N., Langematz, U., and Shibata, K.: On the representation of major stratospheric warmings in reanalyses, Atmospheric Chemistry and Physics, 19, 9469–9484, 2019.

Baldwin, M. P. and Thompson, D. W.: A critical comparison of stratosphere–troposphere coupling indices, Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 135, 1661–1672, 2009.

Baldwin, M. P., Ayarzagüena, B., Birner, T., Butchart, N., Butler, A. H., Charlton-Perez, A. J., Domeisen, D. I., Garfinkel, C. I., Garny, H.,
580 Gerber, E. P., et al.: Sudden stratospheric warmings, Reviews of Geophysics, 59, e2020RG000 708, 2021.

Bayr, T., Domeisen, D. I., and Wengel, C.: The effect of the equatorial Pacific cold SST bias on simulated ENSO teleconnections to the North Pacific and California, Climate Dynamics, 53, 3771–3789, 2019.

Bergner, N., Friedel, M., Domeisen, D. I., Waugh, D., and Chiodo, G.: Exploring the link between austral stratospheric polar vortex anomalies and surface climate in chemistry-climate models, Atmospheric Chemistry and Physics, 22, 13 915–13 934, 2022.

585 Bland, J., Gray, S., Methven, J., and Forbes, R.: Characterising extratropical near-tropopause analysis humidity biases and their radiative effects on temperature forecasts, Quarterly Journal of the Royal Meteorological Society, 147, 3878–3898, https://doi.org/https://doi.org/10.1002/qj.4150, 2021.

Boljka, L. and Birner, T.: Tropopause-level planetary wave source and its role in two-way troposphere–stratosphere coupling, Weather and Climate Dynamics, 1, 555–575, 2020.

590 Boville, B. A.: The influence of the polar night jet on the tropospheric circulation in a GCM, Journal of Atmospheric Sciences, 41, 1132–1142, 1984.

Butler, A., Charlton-Perez, A., Domeisen, D. I. V., Garfinkel, C., Gerber, E. P., Hitchcock, P., Karpechko, A. Y., Maycock, A. C., Sigmond, M., Simpson, I., and Son, S.-W.: Chapter 11 - Sub-seasonal Predictability and the Stratosphere, in: Sub-Seasonal to Seasonal Prediction, edited by Robertson, A. W. and Vitart, F., pp. 223–241, Elsevier, https://doi.org/10.1016/B978-0-12-811714-9.00011-5, 2019.

595 Butler, A. H., Arribas, A., Athanassiadou, M., Baehr, J., Calvo, N., Charlton-Perez, A., Déqué, M., Domeisen, D. I., Fröhlich, K., Hendon, H., et al.: The climate-system historical forecast project: Do stratosphere-resolving models make better seasonal climate predictions in boreal winter?, Quarterly Journal of the Royal Meteorological Society, 142, 1413–1427, 2016.

Byrne, N. J. and Shepherd, T. G.: Seasonal persistence of circulation anomalies in the Southern Hemisphere stratosphere and its implications for the troposphere, Journal of Climate, 31, 3467–3483, 2018.

600 Charlesworth, E., Plöger, F., Birner, T., Baikhadzhaev, R., Abalos, M., Abraham, N. L., Akiyoshi, H., Bekki, S., Dennison, F., Jöckel, P., et al.: Stratospheric water vapor affecting atmospheric circulation, Nature Communications, 14, 3925, 2023.

Charlton-Perez, A. J., Baldwin, M. P., Birner, T., Black, R. X., Butler, A. H., Calvo, N., Davis, N. A., Gerber, E. P., Gillett, N., Hardiman, S., Kim, J., Krüger, K., Lee, Y.-Y., Manzini, E., McDaniel, B. A., Polvani, L., Reichler, T., Shaw, T. A., Sigmond, M., Son, S.-W., Toohey, M., Wilcox, L., Yoden, S., Christiansen, B., Lott, F., Shindell, D., Yukimoto, S., and Watanabe, S.: On the Lack of Stratospheric
605 Dynamical Variability in Low-Top Versions of the CMIP5 Models, Journal of Geophysical Research: Atmospheres, 118, 2494–2505, https://doi.org/10.1002/jgrd.50125, 2013.

Charney, J. G. and Drazin, P. G.: Propagation of planetary-scale disturbances from the lower into the upper atmosphere, Journal of Geophysical Research, 66, 83–109, 1961.

Domeisen, D. I., Butler, A. H., Charlton-Perez, A. J., Ayarzagüena, B., Baldwin, M. P., Dunn-Sigouin, E., Furtado, J. C., Garfinkel, C. I., Hitchcock, P., Karpechko, A. Y., Kim, H., Knight, J., Lang, A. L., Lim, E.-P., Marshall, A., Roff, G., Schwartz, C., Simpson, I. R., Son, S.-W., and Taguchi, M.: The Role of the Stratosphere in Subseasonal to Seasonal Prediction: 1. Predictability of the Stratosphere, Journal of Geophysical Research: Atmospheres, 125, e2019JD030 920, https://doi.org/10.1029/2019JD030920, 2020a.

Domeisen, D. I. V. and Butler, A. H.: Stratospheric Drivers of Extreme Events at the Earth's Surface, Commun Earth Environ, 1, 1–8, https://doi.org/10.1038/s43247-020-00060-z, 2020.

Domeisen, D. I. V., Sun, L., and Chen, G.: The role of synoptic eddies in the tropospheric response to stratospheric variability, Geophysical Research Letters, 40, 4933–4937, https://doi.org/10.1002/grl.50943, 2013.

Domeisen, D. I. V., Butler, A. H., Charlton-Perez, A. J., Ayarzagüena, B., Baldwin, M. P., Dunn-Sigouin, E., Furtado, J. C., Garfinkel, C. I., Hitchcock, P., Karpechko, A. Y., Kim, H., Knight, J., Lang, A. L., Lim, E.-P., Marshall, A., Roff, G., Schwartz, C., Simpson, I. R., Son, S.-W., and Taguchi, M.: The Role of the Stratosphere in Subseasonal to Seasonal Prediction: 2. Predictability Arising From Stratosphere-Troposphere Coupling, Journal of Geophysical Research: Atmospheres, 125, e2019JD030 923, https://doi.org/10.1029/2019JD030923, 2020b.

Dunn-Sigouin, E. and Shaw, T. A.: Comparing and contrasting extreme stratospheric events, including their coupling to the tropospheric circulation, Journal of Geophysical Research: Atmospheres, 120, 1374–1390, 2015.

Eyring, V., Butchart, N., Waugh, D. W., Akiyoshi, H., Austin, J., Bekki, S., Bodeker, G. E., Boville, B. A., Brühl, C., Chipperfield, M. P., Cordero, E., Dameris, M., Deushi, M., Fioletov, V. E., Frith, S. M., Garcia, R. R., Gettelman, A., Giorgetta, M. A., Grewe, V., Jourdain, L., Kinnison, D. E., Mancini, E., Manzini, E., Marchand, M., Marsh, D. R., Nagashima, T., Newman, P. A., Nielsen, J. E., Pawson, S., Pitari, G., Plummer, D. A., Rozanov, E., Schraner, M., Shepherd, T. G., Shibata, K., Stolarski, R. S., Struthers, H., Tian, W., and Yoshiki, M.: Assessment of temperature, trace species, and ozone in chemistry-climate model simulations of the recent past, Journal of Geophysical Research: Atmospheres, 111, https://doi.org/https://doi.org/10.1029/2006JD007327, 2006.

Feldstein, S. and Lee, S.: Is the Atmospheric Zonal Index Driven by an Eddy Feedback?, Journal of the Atmospheric Sciences, 55, 3077 – 3086, https://doi.org/10.1175/1520-0469(1998)055<3077:ITAZID>2.0.CO;2, 1998.

Friedel, M., Chiodo, G., Stenke, A., Domeisen, D. I., Fueglistaler, S., Anet, J. G., and Peter, T.: Springtime arctic ozone depletion forces northern hemisphere climate anomalies, Nature Geoscience, 15, 541–547, 2022a.

Friedel, M., Chiodo, G., Stenke, A., Domeisen, D. I., and Peter, T.: Effects of Arctic ozone on the stratospheric spring onset and its surface impact, Atmospheric Chemistry and Physics, 22, 13 997–14 017, 2022b.

Fujiwara, M., Manney, G. L., Gray, L. J., and Wright, J. S.: SPARC Reanalysis Intercomparison Project (S-RIP) Final Report, SPARC Report No. 10, https://doi.org/10.17874/800dee57d13, 2021.

Garfinkel, C. I., Hartmann, D. L., and Sassi, F.: Tropospheric Precursors of Anomalous Northern Hemisphere Stratospheric Polar Vortices, Journal of Climate, 23, 3282–3299, https://doi.org/10.1175/2010JCLI3010.1, 2010.

Garfinkel, C. I., Schwartz, C., White, I. P., and Rao, J.: Predictability of the early winter Arctic oscillation from autumn Eurasian snowcover in subseasonal forecast models, Climate Dynamics, 55, 961–974, 2020.

Garfinkel, C. I., Chen, W., Li, Y., Schwartz, C., Yadav, P., and Domeisen, D.: The winter North Pacific teleconnection in response to ENSO and the MJO in operational subseasonal forecasting models is too weak, Journal of Climate, 35, 8013–8030, 2022.

Garfinkel, C. I., White, I., Gerber, E. P., Son, S.-W., and Jucker, M.: Stationary Waves Weaken and Delay the Near-Surface Response to Stratospheric Ozone Depletion, Journal of Climate, 36, 565–583, https://doi.org/10.1175/JCLI-D-21-0874.1, 2023.

Garfinkel, C. I., Knight, J., Taguchi, M., Schwartz, C., Cohen, J., Chen, W., Butler, A. H., and Domeisen, D. I.: Development of the signal-to-noise paradox in subseasonal forecasting models: When? Where? Why?, Quarterly Journal of the Royal Meteorological Society, https://doi.org/https://doi.org/10.1002/qj.4822, 2024.

Gerber, E. P. and Martineau, P.: Quantifying the Variability of the Annular Modes: Reanalysis Uncertainty vs. Sampling Uncertainty, Atmospheric Chemistry and Physics, 18, 17 099–17 117, https://doi.org/10.5194/acp-18-17099-2018, 2018.

Guan, H., Zhu, Y., Sinsky, E., Fu, B., Li, W., Zhou, X., Xue, X., Hou, D., Peng, J., Nageswararao, M. M., Tallapgragada, V., Hamill, T. M., Whitaker, J. S., Pegion, P., Frederick, S., Rosencrans, M., and Kumar, A.: GEFSv12 Reforecast Dataset for Supporting Subseasonal and Hydrometeorological Applications, submitted to Monthly Weather Review, 2021.

Hamill, T. M., Whitaker, J. S., Shlyaeva, A., Bates, G., Fredrick, S., Pegion, P., Sinsky, E., Zhu, Y., Tallapragada, V., Guan, H., Zhou, X., and Woollen, J.: The Reanalysis for the Global Ensemble Forecast System, Version 12., Monthly Weather Review, -1, https://doi.org/10.1175/MWR-D-21-0023.1, 2021.

Haynes, P., McIntyre, M., Shepherd, T., Marks, C., and Shine, K. P.: On the "downward control" of extratropical diabatic circulations by eddy-induced mean zonal forces, Journal of the Atmospheric Sciences, 48, 651–678, 1991.

Hendon, H., Lim, E.-P., and Abhik, S.: Impact of interannual ozone variations on the downward coupling of the 2002 Southern Hemisphere stratospheric warming, Journal of Geophysical Research: Atmospheres, 125, e2020JD032 952, 2020.

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

Hitchcock, P. and Simpson, I. R.: The downward influence of stratospheric sudden warmings, Journal of the Atmospheric Sciences, 71, 3856–3876, 2014.

Hitchcock, P., Shepherd, T. G., Taguchi, M., Yoden, S., and Noguchi, S.: Lower-stratospheric radiative damping and polar-night jet oscillation events, Journal of the Atmospheric Sciences, 70, 1391–1408, 2013.

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

Kolstad, E. W., Wulff, C. O., Domeisen, D. I. V., and Woollings, T.: Tracing North Atlantic Oscillation Forecast Errors to Stratospheric Origins, Journal of Climate, 33, 9145–9157, https://doi.org/10.1175/JCLI-D-20-0270.1, 2020.

Lawrence, B. N.: Some aspects of the sensitivity of stratospheric climate simulation to model lid height, Journal of Geophysical Research: Atmospheres, 102, 23 805–23 811, 1997.

Lawrence, Z. D., Abalos, M., Ayarzagüena, B., Barriopedro, D., Butler, A. H., Calvo, N., de la Cámara, A., Charlton-Perez, A., Domeisen, D. I., Dunn-Sigouin, E., et al.: Quantifying stratospheric biases and identifying their potential sources in subseasonal forecast systems, Weather and Climate Dynamics Discussions, 2022, 1–37, 2022.

Lawrence, Z. D., Elsbury, D., Butler, A. H., Perlwitz, J., Albers, J. R., Ciasto, L. M., and Ray, E.: Evaluation of Processes Related to Stratosphere–Troposphere Coupling in GEFSv12 Subseasonal Hindcasts, Monthly Weather Review, 151, 1735–1755, 2023.

Lee, R. W. and Charlton-Perez, A. J.: Diversity of Stratospheric Error Growth Across Subseasonal Prediction Systems, Geophysical Research Letters, 51, e2023GL107 574, https://doi.org/https://doi.org/10.1029/2023GL107574, e2023GL107574 2023GL107574, 2024.

680   Lee, S. H., Charlton-Perez, A. J., Furtado, J. C., and Woolnough, S. J.: Representation of the Scandinavia–Greenland pattern and its relationship with the polar vortex in S2S forecast models, Quarterly Journal of the Royal Meteorological Society, 146, 4083–4098, https://doi.org/https://doi.org/10.1002/qj.3892, 2020.

Lim, E.-P., Hendon, H. H., and Thompson, D. W.: Seasonal evolution of stratosphere-troposphere coupling in the Southern Hemisphere and implications for the predictability of surface climate, Journal of Geophysical Research: Atmospheres, 123, 12–002, 2018.

685   Limpasuvan, V., Thompson, D. W., and Hartmann, D. L.: The life cycle of the Northern Hemisphere sudden stratospheric warmings, Journal of climate, 17, 2584–2596, 2004.

Long, C. S., Fujiwara, M., Davis, S., Mitchell, D. M., and Wright, C. J.: Climatology and Interannual Variability of Dynamic Variables in Multiple Reanalyses Evaluated by the SPARC Reanalysis Intercomparison Project (S-RIP), Atmospheric Chemistry and Physics, 17, 14 593–14 629, https://doi.org/10.5194/acp-17-14593-2017, 2017.

690   Lorenz, D. J. and Hartmann, D. L.: Eddy–zonal flow feedback in the Northern Hemisphere winter, Journal of climate, 16, 1212–1227, 2003.

Lott, F., Rani, R., McLandress, C., Podglajen, A., Bushell, A., Bramberger, M., Lee, H.-K., Alexander, J., Anstey, J., Chun, H.-Y., et al.: Comparison between non orographic gravity wave parameterizations used in QBOi models and Strateole 2 constant level balloons, Quarterly Journal of the Royal Meteorological Society, 2024.

Marshall, A. G. and Scaife, A. A.: Improved predictability of stratospheric sudden warming events in an atmospheric
695   general circulation model with enhanced stratospheric resolution, Journal of Geophysical Research: Atmospheres, 115, https://doi.org/https://doi.org/10.1029/2009JD012643, 2010.

Matsuno, T.: Vertical propagation of stationary planetary waves in the winter Northern Hemisphere, Journal of Atmospheric Sciences, 27, 871–883, 1970.

Oh, J., Son, S.-W., Choi, J., Lim, E.-P., Garfinkel, C., Hendon, H., Kim, Y., and Kang, H.-S.: Impact of stratospheric ozone on the subseasonal
700   prediction in the southern hemisphere spring, Progress in Earth and Planetary Science, 9, 25, 2022.

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

Richter, J. H., Glanville, A. A., Edwards, J., Kauffman, B., Davis, N. A., Jaye, A., Kim, H., Pedatella, N. M., Sun, L., Berner, J., Kim, W. M., Yeager, S. G., Danabasoglu, G., Caron, J. M., and Oleson, K. W.: Subseasonal Earth System Prediction with CESM2, Weather and
705   Forecasting, 37, 797 – 815, https://doi.org/10.1175/WAF-D-21-0163.1, 2022.

Riese, M., Ploeger, F., Rap, A., Vogel, B., Konopka, P., Dameris, M., and Forster, P.: Impact of uncertainties in atmospheric mixing on simulated UTLS composition and related radiative effects, Journal of Geophysical Research: Atmospheres, 117, 2012.

Romanowsky, E., Handorf, D., Jaiser, R., Wohltmann, I., Dorn, W., Ukita, J., Cohen, J., Dethloff, K., and Rex, M.: The role of stratospheric ozone for Arctic-midlatitude linkages, Scientific Reports, 9, 7962, 2019.

710   Scaife, A. A., Baldwin, M. P., Butler, A. H., Charlton-Perez, A. J., Domeisen, D. I., Garfinkel, C. I., Hardiman, S. C., Haynes, P., Karpechko, A. Y., Lim, E.-P., et al.: Long-range prediction and the stratosphere, Atmospheric chemistry and physics, 22, 2601–2623, 2022.

Schwartz, C. and Garfinkel, C. I.: Troposphere-stratosphere coupling in subseasonal-to-seasonal models and its importance for a realistic extratropical response to the Madden-Julian Oscillation, Journal of Geophysical Research: Atmospheres, 125, e2019JD032 043, 2020.

Schwartz, C., Garfinkel, C. I., Yadav, P., Chen, W., and Domeisen, D. I. V.: Stationary wave biases and their effect on upward troposphere–
715   stratosphere coupling in sub-seasonal prediction models, Weather and Climate Dynamics, 3, 679–692, https://doi.org/10.5194/wcd-3-679-2022, 2022.

Scinocca, J. and Haynes, P.: Dynamical forcing of stratospheric planetary waves by tropospheric baroclinic eddies, Journal of the Atmospheric Sciences, 55, 2361–2392, 1998.

Shaw, T. A. and Perlwitz, J.: The Impact of Stratospheric Model Configuration on Planetary-Scale Waves in Northern Hemisphere Winter, Journal of Climate, 23, 3369–3389, https://doi.org/10.1175/2010JCLI3438.1, 2010.

Simpson, I., Hitchcock, P., Shepherd, T., and Scinocca, J.: Stratospheric variability and tropospheric annular-mode timescales, Geophysical Research Letters, 38, 2011.

Simpson, I. R., Shepherd, T. G., Hitchcock, P., and Scinocca, J. F.: Southern Annular Mode Dynamics in Observations and Models. Part II: Eddy Feedbacks, Journal of Climate, 26, 5220 – 5241, https://doi.org/10.1175/JCLI-D-12-00495.1, 2013.

Smith, K. L. and Kushner, P. J.: Linear interference and the initiation of extratropical stratosphere-troposphere interactions, Journal of Geophysical Research: Atmospheres, 117, 2012.

Song, Y. and Robinson, W. A.: Dynamical Mechanisms for Stratospheric Influences on the Troposphere, Journal of the Atmospheric Sciences, 61, 1711 – 1725, https://doi.org/10.1175/1520-0469(2004)061<1711:DMFSIO>2.0.CO;2, 2004.

Stan, C., Zheng, C., Chang, E. K.-M., Domeisen, D. I., Garfinkel, C. I., Jenney, A. M., Kim, H., Lim, Y.-K., Lin, H., Robertson, A., et al.: Advances in the prediction of MJO teleconnections in the S2S forecast systems, Bulletin of the American Meteorological Society, 103, E1426–E1447, 2022.

Taguchi, M.: Comparison of Subseasonal-to-Seasonal Model Forecasts for Major Stratospheric Sudden Warmings, Journal of Geophysical Research: Atmospheres, 123, 10 231–10 247, https://doi.org/10.1029/2018JD028755, 2018.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., Hendon, H., Hodgson, J., Kang, H.-S., Kumar, A., Lin, H., Liu, G., Liu, X., Malguzzi, P., Mallas, I., Manoussakis, M., Mastrangelo, D., MacLachlan, C., McLean, P., Minami, A., Mladek, R., Nakazawa, T., Najm, S., Nie, Y., Rixen, M., Robertson, A. W., Ruti, P., Sun, C., Takaya, Y., Tolstykh, M., Venuti, F., Waliser, D., Woolnough, S., Wu, T., Won, D.-J., Xiao, H., Zaripov, R., and Zhang, L.: The Subseasonal to Seasonal (S2S) Prediction Project Database, Bulletin of the American Meteorological Society, 98, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1, 2017.

Weinberger, I., Garfinkel, C. I., Harnik, N., and Paldor, N.: Transmission and Reflection of Upward-Propagating Rossby Waves in the Lowermost Stratosphere: Importance of the Tropopause Inversion Layer, Journal of the Atmospheric Sciences, 79, 3263–3274, 2022.

White, I. P., Garfinkel, C. I., Gerber, E. P., Jucker, M., Hitchcock, P., and Rao, J.: The generic nature of the tropospheric response to sudden stratospheric warmings, Journal of Climate, 33, 5589–5610, 2020.

White, I. P., Garfinkel, C. I., and Hitchcock, P.: On the tropospheric response to transient stratospheric momentum torques, Journal of the Atmospheric Sciences, 79, 2041–2058, 2022.

Wicker, W., Polichtchouk, I., and Domeisen, D. I.: Increased vertical resolution in the stratosphere reveals role of gravity waves after sudden stratospheric warmings, Weather and Climate Dynamics, 4, 81–93, 2023.

Williams, N. C., Scaife, A. A., and Screen, J. A.: Underpredicted ENSO teleconnections in seasonal forecasts, Geophysical Research Letters, 50, e2022GL101 689, 2023.