



Linking stratosphere–troposphere coupling biases to Annular Mode predictability and midlatitude impacts in subseasonal forecast systems

Chaim I. Garfinkel¹, Zachary D. Lawrence^{2,3}, Amy H. Butler⁴, Blanca Ayarzagüena^{5,6}, Marta Abalos⁵, David Barriopedro⁶, Natalia Calvo^{5,6}, Alvaro de la Cámara⁵, Andrew Charlton-Perez⁷, Judah Cohen⁸, Ying Dai⁹, Daniela I.V. Domeisen¹⁰, Etienne Dunn-Sigouin¹¹, Irene Erner¹², Javier García-Serrano¹³, Neil P. Hindley¹⁴, Peter Hitchcock⁹, Martin Jucker¹⁵, Alexey Yu. Karpechko¹², Hera Kim¹⁶, Gerbrand Koren¹⁷, Simon H. Lee¹⁸, Robert W. Lee⁷, Marisol Osman^{19,20}, Froila M. Palmeiro¹³, Jian Rao²¹, Chen Schwartz^{1,22}, Seok-Woo Son¹⁶, Corwin J. Wright¹⁴, and Rachel W.-Y. Wu²³

¹Fredy & Nadine Herrmann Institute of Earth Sciences, The Hebrew University of Jerusalem, Israel

²Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder, Boulder, CO, USA

³NOAA Physical Sciences Laboratory (PSL), Boulder, CO, USA

⁴NOAA Chemical Sciences Laboratory (CSL), Boulder, CO, USA

⁵Department of Earth Physics and Astrophysics, Universidad Complutense de Madrid, Madrid, Spain

⁶Instituto de Geociencias (IGEO), Consejo Superior de Investigaciones Científicas – Universidad Complutense de Madrid (CSIC-UCM), Madrid, Spain.

⁷Department of Meteorology, University of Reading, Reading, UK

⁸Atmospheric and Environmental Research Inc., Lexington, MA, USA.

⁹Dept. Earth and Atmospheric Sciences, Cornell University, Ithaca, NY, USA

¹⁰University of Lausanne, Lausanne, Switzerland

¹¹NORCE Norwegian Research Centre and Bjerknes Centre for Climate Research, Bergen, Norway

¹²Finnish Meteorological Institute, Meteorological Research, Helsinki, Finland

¹³Group of Meteorology, Universitat de Barcelona (UB), Barcelona, Spain

¹⁴Centre for Climate Adaptation and Environment Research, University of Bath, Bath, UK

¹⁵Climate Change Research Centre and ARC Centre of Excellence for the Weather of the 21st Century, University of New South Wales, Sydney, Australia

¹⁶School of Earth and Environmental Sciences, Seoul National University, South Korea

¹⁷Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, The Netherlands

¹⁸School of Earth and Environmental Sciences, University of St Andrews, St Andrews, UK

¹⁹CONICET – Universidad de Buenos Aires, Centro de Investigaciones del Mar y la Atmósfera (CIMA), Buenos Aires, Argentina

²⁰CNRS – IRD – CONICET – UBA. Instituto Franco-Argentino para el Estudio del Clima y sus Impactos (IRL 3351 IFAECI), Buenos Aires, Argentina

²¹Collaborative 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

²²Centre for Climate Research Singapore, Singapore

²³Department of Applied Physics and Applied Mathematics, Columbia University, New York, USA

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

Abstract.



Two-way coupling between the stratosphere and troposphere is an important source of subseasonal predictability and can open windows of opportunity for improved forecasts. However, operational models used for subseasonal forecasting suffer from systematic biases due to insufficient resolution or poorly performing parameterizations: they underestimate the strength of upward coupling from the troposphere to the stratosphere, they overestimate downward coupling of the lower stratosphere to near-surface polar cap height, and the polar lower stratosphere is too cold. Using data from 22 operational subseasonal models, we explore the implications of these biases on the skill of the Northern and Southern Annular Modes (NAM and SAM) in both the troposphere and stratosphere and for regional surface impacts. Updated or upgraded model versions from nearly all modeling centers show higher skill in both the stratospheric and tropospheric NAM and SAM as compared to their earlier versions, and models with relatively higher skill for the stratospheric annular modes generally have higher skill for the tropospheric annular modes as well. Near-surface AM skill is significantly higher for initializations during strong vortex states, and to a lesser degree during weak vortex states. Models with particularly poor upward coupling strengths and with too-weak climatological heat flux at 100 hPa suffer from reduced stratospheric NAM and SAM skill. In contrast, mean state biases in the stratospheric polar vortex are relatively unimportant for stratospheric NAM and SAM skill. Even as downward coupling over the North Atlantic and Antarctica tends to be handled well by these models, they produce an overly zonal response with a fictitious center of action in the midlatitude North Pacific, while underestimating the midlatitude near-surface temperature impacts over Eurasia, Australia, and South America. Downward coupling in the North Pacific is particularly pronounced in models with overly strong variance in this region, and models underestimate the wave-3 pattern in Southern Hemisphere (SH) midlatitudes by a factor of three. These biases in downward coupling mirror biases in the midlatitude structure of the tropospheric NAM and SAM. Overall, models with finer vertical resolution near the tropopause, with a better representation of tropospheric quasi-stationary waves, and with a better representation of these coupling processes, are more skillful in both the stratosphere and troposphere. These findings have implications for future research directions of the Stratospheric Network for the Assessment of Predictability.

1 Introduction

The Southern and Northern Annular Modes (SAM and NAM) are the dominant modes of variability in the extratropical atmosphere, describing large-scale fluctuations in pressure between the polar regions and the midlatitudes (Thompson and Wallace, 2000). The SAM is largely zonally symmetric, whereas the NAM is not and is dominated by variability in the North Atlantic sector, similar to the North Atlantic Oscillation (NAO; Thompson and Wallace, 1998). When the NAM is in its positive phase, lower pressure over the Arctic and poleward shifted westerly winds tend to confine cold air to higher latitudes, often leading to milder winter conditions in regions like Europe and parts of North America (Thompson and Wallace, 2001). In its negative phase, the near-surface jet and storm tracks typically shift equatorward, increasing the likelihood of severe winter weather in northern Europe and extreme precipitation in southern Europe (Thompson and Wallace, 1998). The SAM also modulates surface climate over Australia, Southern Africa, and South America (Gillett et al., 2006); for example, the negative phase of the SAM often leads to hot and dry weather in southeastern Australia, and the positive phase to enhanced rainfall



35 in southeastern South America (Purich et al., 2026). Through their association with local weather, the SAM and NAM affect agriculture, energy demand, transportation, and extreme weather risk (Thompson and Wallace, 2001; Fogt and Marshall, 2020). Hence, accurately predicting them can lead to improved mitigation of climate-related impacts and weather extremes.

The tropospheric annular modes are year-round phenomena, but the NAM has a longer characteristic timescale in the boreal extended winter and the the SAM has a longer characteristic timescale in the austral spring and summer (Gerber
 40 et al., 2008, 2010). This longer characteristic timescale allows for enhanced predictability in these seasons (Son et al., 2020; Domeisen et al., 2020a), and thus the potential for windows of opportunity for subseasonal prediction (Butler et al., 2026; Domeisen et al., 2020b). These longer timescales are linked to coupling between variability in the polar stratosphere and the surface annular modes (Thompson et al., 2003; Simpson et al., 2011), and are associated with long radiative timescales in the lower stratosphere (Hitchcock et al., 2013) that allow for polar stratospheric vortex extremes to influence surface climate and
 45 weather extremes for weeks to months afterwards (Domeisen and Butler, 2020).

The vertical coupling is mediated by dynamical interactions between planetary-scale Rossby waves and the mean flow, and operates in both directions: upward coupling from tropospheric variability modulates the strength of the stratospheric polar vortex, while downward coupling from polar stratospheric variability can influence the surface annular modes and surface weather (Scaife et al., 2022; Butler et al., 2026; Hong and Son, 2025). Model biases in either the troposphere or stratosphere
 50 can degrade these coupling processes, and thereby compromise opportunities to increase subseasonal predictability that could otherwise be achieved. 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 have previously diagnosed biases in the extratropical atmospheric mean state (Lawrence et al., 2022) and in the representation of stratosphere–troposphere coupling processes (Garfinkel et al., 2025) in subseasonal forecast systems. Lawrence et al. (2022)
 55 and Garfinkel et al. (2025) identified several systematic model biases of relevance to the NAM and SAM. Most forecast systems exhibit a warm bias in the global-mean stratosphere, a cold bias in the extratropical lower stratosphere-upper troposphere, and a strong and cold polar vortex bias. Furthermore, they underestimate the strength of upward coupling from the troposphere to the stratosphere in the Northern Hemisphere (NH), while they overestimate downward coupling of the lower stratosphere to near-surface polar cap height. Nearly all of these biases are substantially worse for models with a low model lid height, and
 60 biases in coupling processes are typically worse in models with larger mean-state biases (Garfinkel et al., 2025). The goal of this study is to evaluate the connection between these systematic biases across subseasonal forecast systems and forecast skill for the NAM and SAM, and also to consider the regional impacts of biases in downward coupling.

Our focus on regional impacts of biases in downward coupling has two primary motivations. First, for many societal applications the primary regions of interest are the midlatitudes, and a biased downward coupling in the polar cap is not inherently
 65 a cause for concern for most end-users. Second, previous work focusing on climate models (Dai et al., 2024) and on seasonal forecasting models (Lee and Polvani, 2024) has found a too-strong center of action of the NAM in the Pacific sector, such that the NAM is overly zonal. The copious daily data available in the subseasonal hindcasts allows for a re-examination of the development of this bias, and of the midlatitude impacts from the polar stratosphere more generally.



After introducing the data and methods in section 2 and documenting the skill at different levels and in multiple model
 70 versions from most centers (Section 3), we demonstrate that models with higher skill at predicting the stratospheric NAM tend
 to have a better representation of the mean state and of coupling processes, and also typically have higher vertical resolution
 (Section 4). We then demonstrate that models systematically misrepresent the regional structure of downward coupling in both
 hemispheres, and specifically simulate too-weak and overly-zonal impacts over midlatitudes even as the downward coupling
 over the polar cap is too strong in the NH (Section 5). We summarize our results and place them in the context of previous
 75 work in Section 6.

2 Data & Methods

2.1 Subseasonal Hindcast and Reanalysis Datasets

We use ensemble hindcast data from the Subseasonal-to-Seasonal (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 Atmo-
 80 spheric Administration’s Global Ensemble Forecast System version 12 (NOAA GEFSv12; Hamill et al., 2022; Guan et al.,
 2022), (ii) the National Center for Atmospheric Research Community Earth System Model version 2 (NCAR-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 compo-
 85 nent (CESM2-WACCM6, hereafter CESM2-WACCM; Richter et al., 2022). Daily gridded latitude-longitude data were only
 retained for 18 forecast systems due to the large data volume, and so metrics which rely on these data are only computed for
 these 18 systems.

The specific model versions and the period used for each system are included in Table 1. The models differ in their treatment
 of chemistry: WACCM has fully interactive chemistry; ECMWF (both cycles), CNRM (both versions), ECCC (GEPS 7 only)
 and NCEP (both versions) use prognostic linearized ozone. All the others prescribe composition to the best of our knowledge.
 90 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 the same year. We therefore have elected not to focus on a common
 period for this analysis.

We compare the hindcast fields to those from the European Centre for Medium-range Weather Forecasts (ECMWF) ERA5/5.1
 95 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.

For each variable and forecast system, we derive lead-time-dependent climatologies ($\overline{F}$), which we subtract from the raw
 100 forecast quantities to determine forecast anomalies. These climatologies are calculated by averaging all ensemble-mean hind-
 casts for a given day of year and for each lead time. For those systems providing a fixed set of hindcast initializations that do



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

Modeling Center	Model Version(s)	Hindcast Period	Initis per season	Ensembles	Forecast Span	Model Top
BoM*	POAMA P24	1999-2010	15	33	62 days	10 hPa
NCAR*	CESM2-CAM	1999-2019	12-13	11	45 days	2 hPa
NCAR	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*	GEPS 4	1995-2014	25-26	4	32 days	2 hPa
ECCC	GEPS 6	1998-2017	23-25	4	32 days	0.1 hPa
ECCC	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
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
NCEP	GEFSv12	2000-2019	12-13	11	35 days	0.1 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.

not uniformly cover the same days of year in the hindcasts (e.g., every Monday and Thursday), we permit differences of up to three days when creating the lead-time-dependent climatologies. The reference climatology (indicated by $\bar{O}$) is based on subsampling ERA5 to match the data availability for each model.

105 2.2 Methods

We define the NAM and SAM using standardized polar cap height anomalies ($60 - 90^\circ$) at 10 hPa, 100 hPa, 850 hPa, and 1000 hPa (Baldwin and Thompson, 2009), which avoids problems with physically interpreting empirical orthogonal function-based patterns in models (Lee and Polvani, 2024). We evaluate skill of the forecast (F) with respect to ERA5 (O for “observations”) using the anomaly correlation coefficient (ACC):



$$110 \quad ACC = \frac{\sum_{i=1}^N (F_i - \bar{F}) (O_i - \bar{O})}{\sqrt{\sum_{i=1}^N (F_i - \bar{F})^2 \sum_{i=1}^N (O_i - \bar{O})^2}} \quad (1)$$

where i refers to each available initialization in December-February (DJF; for the NH) or September-November (SON; for the SH) over the hindcast period. We primarily evaluate skill in days 22–28 (week 4) after initialization and downward coupling in week 4 or later, and in all analyses we subsample reanalysis data to match each system. This averaging over week 4 is performed before we compute the skill, and we compute skill using the ensemble mean of each initialization. Because
 115 our focus is on the latter half of the available lead time, the verification window for skill and coupling lies primarily in late September to mid December for the SH and late December to mid March for the NH. This choice covers the dynamically active seasons in both hemispheres; note that most subseasonal models suffer from a too-early SH vortex breakdown, opposite to the bias in many climate models (Lawrence et al., 2022). There is no exact overlap in the analysis period across models, and so the precise verification window differs across models. We therefore cannot rule out the possibility of genuine interannual
 120 variability in the predictability of the NAM/SAM that may alias into our assessment of the skill of each model. More generally, we are using the available hindcasts as an ensemble of opportunity, and so it is impossible to constrain every possible factor that could affect skill.

We then compare intermodel spread in the ACC to intermodel spread in vertical resolution, in biases in the mean state, and in biases in coupling processes. We focus on biases in the following eight key metrics to diagnose coupling strength throughout
 125 this paper: meridional eddy heat flux $\overline{v'_{k=1}T'_{k=1}}$ and $\overline{v'_{k=2}T'_{k=2}}$ averaged over 45–75° in each hemisphere [500 and 100 hPa]; polar cap (60 – 90°) geopotential heights [10, 100, 850, and 1000 hPa] and temperature [100 hPa]. v denotes the meridional wind, T the temperature, a horizontal bar over a quantity indicates the zonal mean and a prime denotes a departure from 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, which is proportional
 130 to the vertical component of the Eliassen-Palm flux and wave group velocity according to linear theory (Edmon Jr et al., 1980). The definition of biases in coupling processes follows Garfinkel et al. (2025), and specifically, we focus on the following metrics:

1. Vertical propagation of planetary waves from the troposphere into the stratosphere, defined as the regression coefficient between 500 hPa and 100 hPa of $\overline{v'T'}$ for both wave-1 and wave-2, with $\overline{v'T'}$ at 100 hPa lagged 3 days, and $\overline{v'T'}$ at
 135 500 hPa is taken from days 11 to 22 after initialization.
2. The sensitivity of the stratospheric polar vortex to upward wave driving from the lowermost stratosphere, defined as the regression coefficient with 10-hPa NAM (NAM10) lagged 7 days after $\overline{v'T'_{k=1+2}}$ at 100hPa for days 11 to 22 after initialization.
3. Downward propagation of stratospheric polar vortex anomalies from the upper- and mid-stratosphere to the lower strato-
 140 sphere related to wave-mean flow interactions, defined as the regression coefficient of NAM10 in days 9 to 12 after initialization with 100-hPa NAM (NAM100) lagged 2 days later.



4. The persistence of the polar vortex signal in the lower stratosphere that arises due to the long radiative timescales and high static stability, defined as the regression coefficient of polar-cap averaged temperature at 100hPa (Tcap100) with itself lagged 20 days later for days 9 to 12 after initialization.

145 5. Downward propagation from the lower stratosphere to the near-surface, defined as the regression coefficient of NAM100 in days 9 to 12 after initialization with NAM850 lagged 5 or 10 days later.

Analogous indices but for the SH are used to quantify stratosphere-troposphere coupling in austral spring. These metrics are computed for each ensemble member separately, and then averaged across members. Additional details on the calculation of these diagnostics and a sensitivity analysis for the precise lags chosen can be found in Garfinkel et al. (2025). For metrics
 150 involving downward coupling, we focus on stratospheric variability for days 9 to 12 only because we then examine surface impacts up to 20 days later, and some of the models only forecast out to day 32. The primary focus is linking skill of the SAM/NAM to the error magnitude (absolute bias), rather than the sign of the error.

For vertical resolution, we focus on the number of model levels from the surface to 300 hPa, from 300 hPa to 100 hPa, and from 100 hPa to 10 hPa. This allows for the exploration of the importance of better resolving the troposphere, the tropopause
 155 and lowermost stratosphere, and the lower- to mid-stratosphere, respectively. Nonetheless, there is a high correlation in the number of levels in each of these layers; for example, the correlation of the number of levels below 300 hPa and the number of levels from 100 to 10 hPa is 0.67. Hence we cannot fully isolate the relative roles of the vertical resolution in each layer due to the strong collinearity.

As in Lawrence et al. (2022), we define forecast systems with model tops at or above 0.1 hPa with several levels above 1
 160 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 in every analysis. We stress that the computation of high-top and low-top model-means is obtained from an unbalanced distribution of high-top and low-top models.

3 NAM/SAM predictability for different initial vortex states, pressure levels, and model generations

165 We begin by demonstrating that the connection between the strength of the lower stratospheric polar vortex (as measured by NAM100) and lower tropospheric NAM skill identified in previous work (e.g., Domeisen et al., 2020b) is also evident in this expanded set of models. We also demonstrate a similar but even stronger stratospheric influence for the lower tropospheric SAM skill. Specifically, Figure 1 shows the weeks 3-4 ensemble-mean skill of the NAM and the SAM at 850 hPa for initializations in which the NAM100/SAM100 exceed the climatological 85th percentile (DJF for the NAM100, SON for the SAM100), fall
 170 below the climatological 15th percentile, or are neutral (fall between the 25th and 75th percentiles). Consistent with previous work (Sigmond et al., 2013; Tripathi et al., 2015; Sun et al., 2020; Domeisen et al., 2020b), for a majority of models and in the multi-model mean (MMM) the NAM is significantly more predictable 3-4 weeks after the lower stratospheric polar vortex is anomalously strong or weak compared to neutral. However, the confidence intervals for individual models are very large and



overlap across different conditional states, likely due to the relatively small number of DJF initializations, the large event-to-
 175 event variability in NH downward coupling, and small ensemble sizes (Karpechko et al., 2025). Moreover, a couple of models
 show higher skill following neutral states, or only higher skill following either the strong or weak states, which may be due to
 the different periods being considered for each model. Using the much shorter “common” period of 1999–2010 produces more
 consistent high skill following initializations with strong NAM100 across models (not shown).

In the SH, the lower tropospheric SAM at weeks 3–4 more consistently shows higher skill for initializations with weak and
 180 strong lower stratospheric vortex strength compared to neutral across models, and the overall week 3–4 SAM skill is higher
 compared to the NAM. The MMM shows roughly 60% higher skill for the SAM following initializations with weak or strong
 states compared to neutral.

To provide context on the skill of each model in predicting the annular modes, we show in Figure 2 the ACC between each
 model and ERA5 for weekly averaged AM at 850, 100, and 10 hPa for the SH and NH. (Note that because these are weekly
 185 averages, the ACC for forecast day 0 is not 1.0). There is a wide range in skill across models, and specifically the ACC for
 AM850 exceeds 0.6 out to day 15 for some models, but only out to day 9 for others. Stratospheric AMs in both hemispheres
 are more skillfully predicted at weeks 3–5 than tropospheric AMs, although the period for which the skill remains above 0.6 is
 more variable across models than for Z850. Skill is higher in the SH than in the NH in both the stratosphere and troposphere,
 and is generally higher for high-top models compared to low-top models. These effects are generally similar to those found in
 190 Domeisen et al. (2020a) and Son et al. (2020); however, our analysis add robustness to previous results as many more high-top
 models have been included here than in these earlier studies.

Next, we consider whether there is a relationship between skill in AMs across the different pressure levels, and additionally
 consider whether there is an improvement in model skill across model generations or model versions from the same center,
 e.g., comparing GloSea5 to GloSea6 or NCEP-CFSv2 to NCEP-GEFSv12. Figure 3 shows the ACC for each model for weekly
 195 averaged AMs at 850, 100, and 10 hPa, and arrows indicate changes in skill across different models for the same modeling
 center, with the arrow pointing towards the more recent or higher resolution model. Consistent with Figure 3 of Domeisen
 et al. (2020a), models with higher skill for AM10 also have higher skill for AM100 and to a lesser degree for AM850. Most,
 though not all, of the updated model versions show higher skill than their older versions. For the NCAR models in particular,
 the added resolution and interactive chemistry (among other differences; Richter et al., 2022) in WACCM versus CAM leads
 200 to a major improvement in SAM10 skill in the SH. In contrast, GloSea6 has lower skill than GloSea5 (both KMA and UKMO)
 for NAM850.

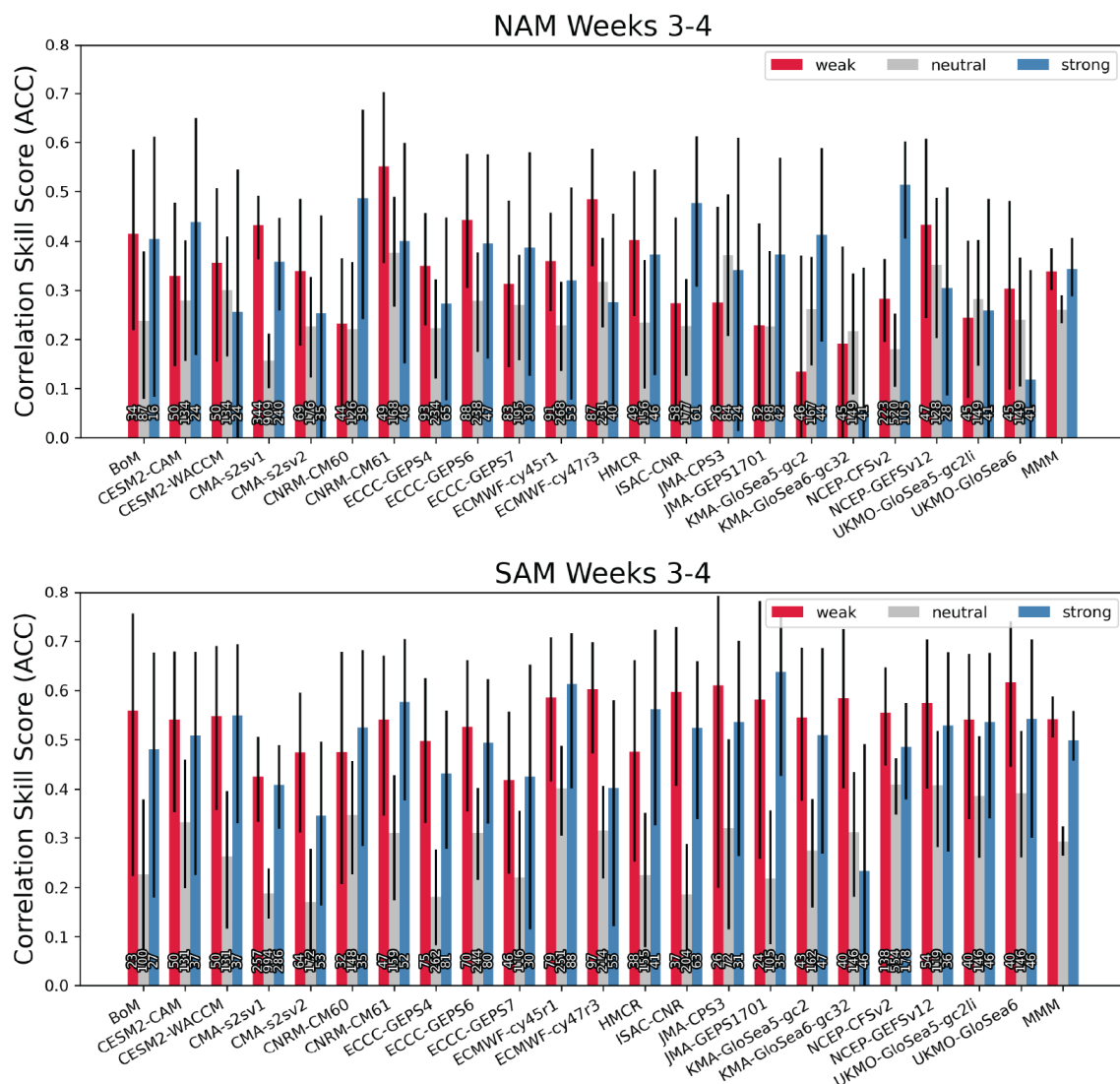


Figure 1. Regime-dependent weeks 3-4 ACC skill for the ensemble-mean NAM and SAM at 850 hPa, conditional on the initial NAM/SAM state at 100 hPa in DJF and SON, respectively. Weak lower stratospheric vortex states (red bars) are defined by NAM/SAM100 less than the 15th percentile; strong lower stratospheric vortex states (blue bars) are defined by NAM/SAM100 greater than the 85th percentile; and neutral lower stratospheric vortex states (gray bars) are defined by NAM/SAM100 between the 25–75th percentile of the DJF or SON 1990-2019 ERA5 climatology. The numbers on the bars indicate the number of initializations for each subsample used in the calculation. Error bars show the 90% bootstrap confidence interval of the correlations for 10000 random samples. The multi-model mean (MMM) is not the skill of the multi-model mean itself, but rather the mean of the individual models' skill. This was calculated using a Fisher-Z transformation of the individual model correlations for each category and then averaging; the bootstrap distributions for each model correlation also had the Fisher-Z transform applied and then averaged to get the multi-model mean uncertainty range.

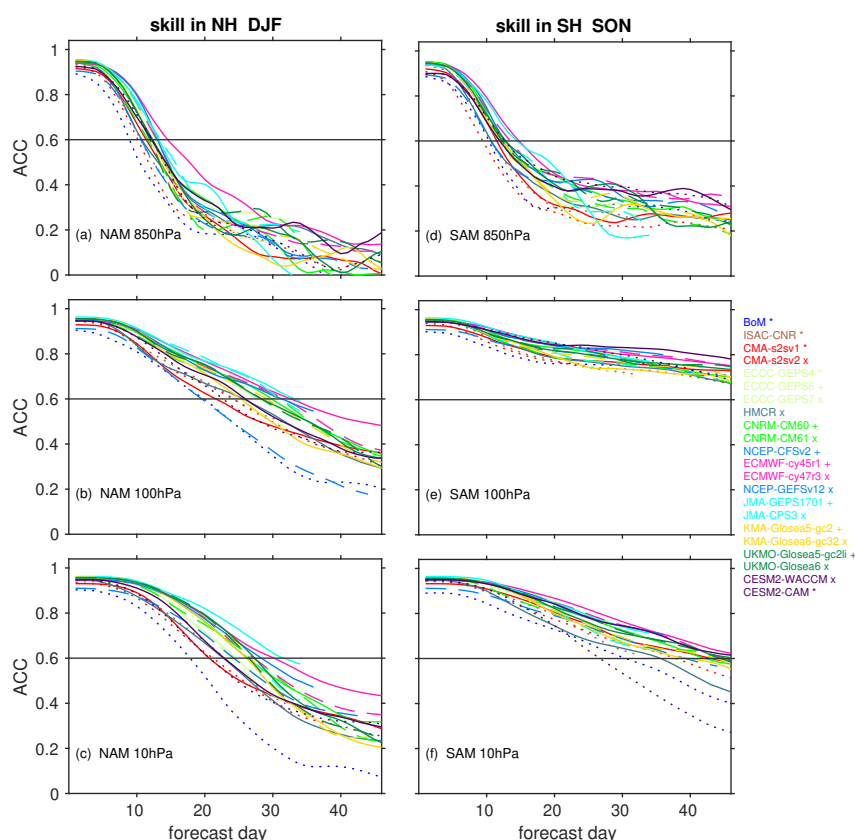


Figure 2. Evolution of ACC of NAM and SAM (Zcap, 60°-pole) at (a, d) 850 hPa, (b, e) 100 hPa, and (c, f) 10 hPa. Low-top models are dotted. Older versions of high-top models are dashed. (a-c) initializations in December-January-February for the NH, (d-f) initializations in September-October-November for the SH. Skill is evaluated after first weekly averaging with the midpoint of the averaging period indicated on the x-axis. For the first three days, the window for the running-mean is truncated and the average is taken over only the available elements (e.g., on day 1, the average is performed on output up to day 4).

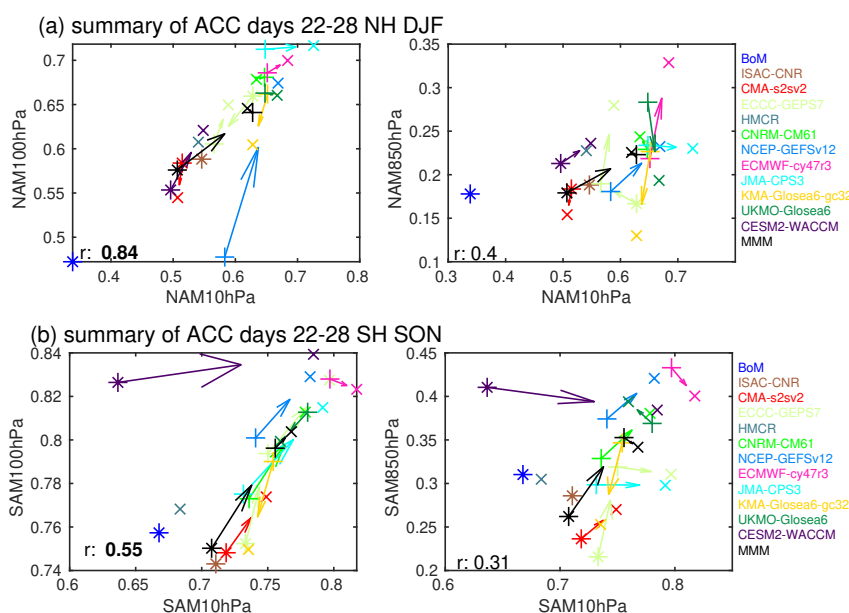


Figure 3. Summary of the ACC for each model (with ERA5 used as reference) in days 22-28 for (a) the NH in DJF and (b) the SH in SON. For different models from the same center, an arrow points towards the more recent version. Low-top models are denoted with an asterisk, older versions of a high-top model with a plus, and the latest version of high-top models with an x. (left) Z10 hPa and Z100 hPa polar cap; (right) Z10 hPa polar cap and Z850 hPa polar cap. The multi-model mean is black. The correlation is indicated in each panel. Bold correlations indicate significant relationships at the 95% confidence level using a two-sided Student's *t*-test.



4 Upward coupling and skill of stratospheric AM

We now turn our attention to the role of vertical resolution and of biases in either the mean state or in coupling processes, for the intermodel spread in skill. Specifically, one might expect that models with poorer performance in coupling strength or in their mean state would have lower skill, while the null hypothesis is that skill and coupling strength are unrelated. Further, Garfinkel et al. (2025) demonstrated that models with poor vertical resolution struggle to capture coupling processes, and this might also lead to poorer skill. We focus on skill of the NAM and SAM at 10 and 850 hPa (AM10 and AM850), and specifically evaluate whether biases linked to upward coupling degrade AM10 predictability in this section, and biases linked to downward coupling degrade AM850 predictability in the next Section (Section 5).

Garfinkel et al. (2025) noted that subseasonal models suffer from a systematic too-weak bias in the strength of upward coupling from the mid-troposphere to the lower stratosphere in the NH. Here, we contrast the magnitude of this upward coupling bias for wave-1 $\overline{v'T'}$ (on the y-axis) with the skill of AM10 (on the x-axis) for the NH and SH in Figures 4a and g, respectively. The same figures but for wave-2 $\overline{v'T'}$ are shown in Figures 4b,h. In both hemispheres, smaller biases in upward wave coupling are significantly correlated with higher skill for AM10. This relationship is evident both when we analyze all 22 models, and also when we focus on high-top models only though generally the effect is weaker. In the NH, this relationship is significant when considering wave-1 $\overline{v'T'}$, while in the SH it is evident for both wave-1 and wave-2 but statistically significant only for wave-2. An additional systematic bias in upward wave coupling identified in Garfinkel et al. (2025) is that the polar vortex is insufficiently sensitive to $\overline{v'T'}$ at 100 hPa in the NH, and is instead overly sensitive in the SH. We consider the relationship of model spread in this bias with the spread in skill in Figures 4c,i. In both hemispheres, a smaller bias in the sensitivity of the polar vortex to wave driving is associated with more skill, but the relationship is much more robust in the NH. Nonetheless, these relationship generally weaken when we exclude the low-top models, and BoM in particular.

In addition to the too-weak bias in upward coupling strength in the NH, Lawrence et al. (2022) and Garfinkel et al. (2025) also identified a too-weak bias in the $\overline{v'T'}$ mean state. In both hemispheres, the models with especially too-weak biases in upward coupling strength also tend to have poor mean states (Figure 6 of Garfinkel et al., 2025). We therefore consider whether mean state biases in $\overline{v'T'}$ are associated with NAM and SAM skill in Figure 5a,b,g,h. Skill of NAM10 in DJF is significantly correlated with mean state biases in wave-1 $\overline{v'T'}$ (Figure 5a), while skill in the SH in SON is significantly correlated with mean state biases in wave-2 $\overline{v'T'}$ (Figure 5h); that is, skill at predicting AM10 is lower for models with a poorer $\overline{v'T'}$ mean state. However, excluding the low-top models, and BoM in particular, weakens the effect. These effects are similar in magnitude to the relationship with upward coupling biases from Figure 4, and hence it is not possible to discern whether a biased mean state or a biased upward coupling strength is more detrimental for skill at predicting AM10. Finally, the relationship between mean state biases in AM10 and skill in predicting AM10 is not robust in either hemisphere (Figure 5c,i).

Lawrence et al. (2022) and Garfinkel et al. (2025) also noted that mean state and upward coupling biases were particularly large in low-top models, though among high-top models the relationship between the number of levels and the magnitude of the biases was not typically statistically significant (Supplemental Figures S9-S11 of Garfinkel et al. (2025)). We assess the role of vertical resolution for the skill of the stratospheric NAM and SAM in Figure 6a-c,g-i. Models with higher vertical

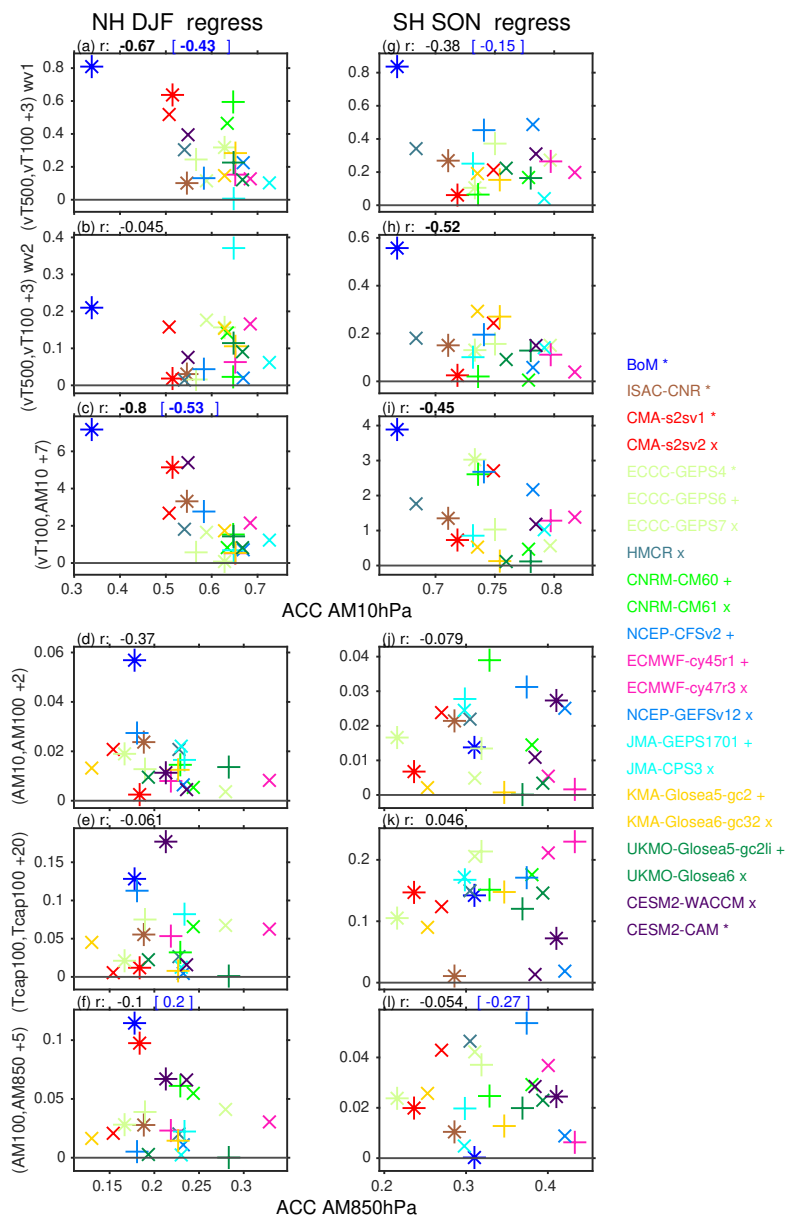


Figure 4. Comparison of the ACC in days 22-28 in AM850 and AM10 to coupling strength biases (see section 2.2). Relationship between (y-axis) the absolute bias in coupling strength as compared to ERA5 for corresponding dates and (x-axis) AM10 ACC (top three rows) AM850 ACC (bottom three rows). Coupling strength is defined using regression; the corresponding figure for correlations is in Supplemental Figure S1. The correlation for each panel is indicated, and also the correlation without low-top models in blue if this correlation differs from the overall correlation by more than 0.2. The left column is for the NH and DJF initializations, and the right column is for the SH and SON initializations. 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 ‘x’. (a,g) upward wave-1 coupling; (b,h) upward wave-2 coupling; (c,i) sensitivity of Z10 hPa polar cap to 100 hPa heat flux; (d,f) coupling strength of Z10 hPa polar cap to Z100 hPa polar cap; (e,k) persistence of T100 hPa polar cap; (f,j) coupling strength of Z100 hPa polar cap to Z850 hPa polar cap. Bold numbers indicate significant relationships at the 95% confidence level using a two-sided Student’s *t*-test.

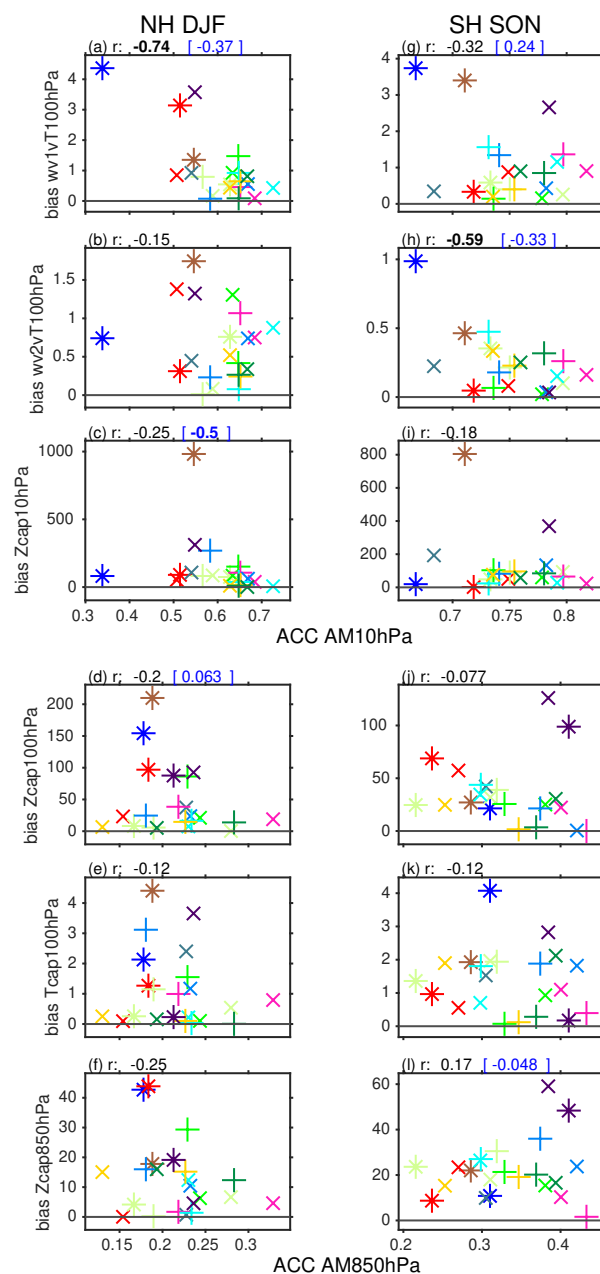


Figure 5. As in Figure 4 but with mean state absolute biases on the y-axis.

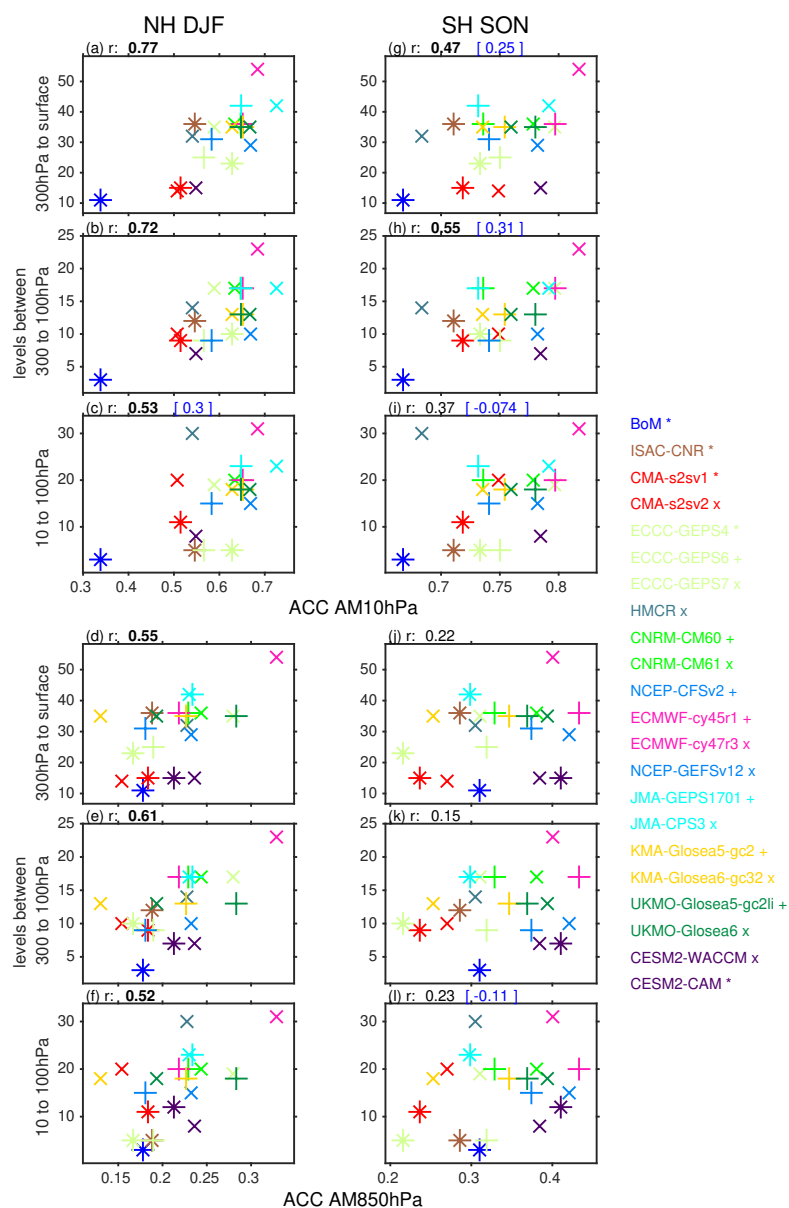


Figure 6. As in Figure 4 but with vertical resolution on the y-axis.



resolution more skillfully predict the stratospheric AM in both hemispheres, with the effect particularly strong for the number of tropospheric or lower stratospheric levels, but less so for stratospheric levels. While one might have naively expected stratospheric resolution to matter more, it becomes clear that a well resolved troposphere and lower stratosphere is crucial: this extra resolution helps with the simulation of the events that generate variability in the $\overline{v'T'}$ that then modulate the vortex strength. This extra resolution can influence where upward propagating waves break or propagate, which can subsequently increase skill higher up and reduce the mean state bias (Wu and Reichler, 2020; Wicker et al., 2023).

Overall, more skillful models tend to have stronger upward coupling between tropospheric $\overline{v'T'}$ and the polar vortex, smaller biases in their $\overline{v'T'}$ mean state, and higher vertical resolution in the troposphere and lower stratosphere. All three of these relationships have quantitatively similar correlation coefficients, and thus with the present experiments it is not possible to distinguish their relative importance or to allow for statements on causality. These relationships are evident even if we exclude low-top models, though they are generally weaker. One possible interpretation is that “the good models are good at everything”, i.e., models that invest in improved vertical resolution may also invest more attention to reducing mean state and coupling biases, and also have better AM10 skill. This is true even as the relationship between resolution and coupling processes is generally not statistically significant (Supplemental Figures S9-S11 of Garfinkel et al., 2025). Results are similar if we focus on lead-lag relationships between biases and skill: the biases evident in days 22-28 are also evident in earlier weeks, including the relative magnitude across models (not shown). Furthermore, the relationship between skill and resolution is stronger in the NH than in the SH, perhaps because skill in the SH is less related to biases in stratosphere–troposphere coupling but rather more strongly influenced by aspects such as biases in chemistry or composition.

5 Downward coupling

5.1 Importance of downward coupling for skill of Annular Modes at 850hPa

We now turn our attention to whether mean state biases or biases in downward coupling are associated with reduced skill for the near-surface NAM and SAM. There is no clear relationship between either mean state biases or biases in downward coupling and the near-surface NAM and SAM skill (Figure 5d,e,f,j,k,l and Figure 4d,e,f,j,k,l). This is true even as skill in the stratospheric NAM/SAM is significantly correlated with skill in the near-surface NAM/SAM across these models (Figure 3), and even as these models show that predictability is enhanced for initializations when the vortex is strong or weak (Figure 1). One possible interpretation of this result is that other effects can overwhelm the role of the stratosphere when evaluated across all initializations. One such effect is resolution, and specifically higher resolution models can better resolve the tropospheric processes that underlie subseasonal prediction. We consider this effect in Figure 6d,e,f,j,k,l, which contrasts the number of model levels with surface AM skill. Additional levels are associated with higher NAM skill, but the strong relationship among models in tropospheric resolution and stratospheric resolution makes it difficult to isolate their relative roles in NAM skill. SAM skill, on the other hand, is not strongly associated with the number of levels, and it is conceivable that spread in SAM skill is dictated by the ability of models to represent other tropospheric subseasonal teleconnections or by the representation of unforced variability.



5.2 Midlatitude tropospheric response to the stratosphere

270 5.2.1 Regional structure of the Northern Annular Mode (NAM)

Downward coupling from the lower stratosphere to the near surface is too strong when evaluated using the polar cap averaged height in most models (Garfinkel et al., 2025). However, for many societal applications the most important region of interest is the midlatitudes, and in particular the anomalies over continents. We now assess whether these impacts are also too strong, and begin by focusing on the temperature impacts over land in Figure 7a-b. This figure shows the correlation of NAM10 with temperature at 850 hPa lagged 10 days later for 18 of the models for which we archived the necessary data. The multi-model mean of the four low-top models with available data is shown in Figure 7a, and of the 14 high-top models in Figure 7b. Supplemental Figure S2 shows each model individually. The models simulate the expected temperature impacts associated with the NAM: a negative NAM leads to cold conditions over northern Europe and Siberia and the eastern United States, and warm conditions over eastern Canada and subtropical Eurasia. The magnitude of these impacts is too weak, however, over 280 subtropical Eurasia in all models, and in subpolar Eurasia in most models and in the multi-model mean. The too-weak bias over subtropical Eurasia exceeds 50% of the magnitude of the observed signal in the multi-model mean. The development of the bias for each model is shown in Figure 8ab, and the bias on day 10 in each individual ensemble member is shown in Figure 9ab. The only model with a bias in subtropical Eurasia smaller than 20% is NCEP-CFSv2 (Figure 8b, Supplemental Figure S2e), and only one individual ensemble member of the 135 available show a correlation as strong as observed (Figure 9b). 285 Biases in downward coupling to subpolar Eurasia tend to be smaller at short lags (Figure 8a) and are not as pervasive across individual members (Figure 9a) or in HMCR, ISAC-CNR, and CNRM-CM60. However, biases over subtropical Eurasia are systematic at all lags (Figure 8b). Results are similar if we consider a regression perspective (Supplemental Figure S3).

This contrast between the polar-cap-averaged NAM response (too strong) and midlatitude temperature impacts (too weak) shows that polar-cap coupling metrics alone are not sufficient diagnostics of impact-relevant downward coupling, and motivates 290 a map view of the Z500 anomalies after anomalous conditions in the polar stratosphere. Figure 7cd shows the correlation of NAM10 with Z500 anomalies 10 days later. Over the polar cap, the downward impact is too strong by 20% (relative to the observed response) in the multi-model mean (top left of Figure 7cd, Garfinkel et al., 2025), with the bias emerging more strongly for later lags (Figure 8c). Nonetheless, in the midlatitude Atlantic, this bias is no longer evident: the Azores center of the NAO is, if anything, slightly too weak in the multi-model mean rather than too strong, with some models suffering from 295 too-weak a center of action (BoM, NCEP-CFSv2, both JMA) and others too-strong a center of action (HMCR, both GloSea6; Supplemental Figure S4). This bias in the midlatitude North Atlantic is also evident when considering Z500 lagged by 5 or 15 days (Supplemental Figure S5 and S6) or when considering a regression perspective (Supplemental Figure S7). Furthermore, the high over Southwest Asia is too weak in nearly all models (CFSv2 and JMA-CP3 notable exceptions). This too-weak response in mid-tropospheric height over the midlatitude Atlantic and southwest Asia is consistent with the too-weak response 300 in subtropical Eurasian temperatures at 850hPa.

There is no robust relationship across models between biases in the strength of the Azores high response and the time-mean strength of the North Atlantic stationary wave ridge (correlation of 0.20; details not shown). Instead, biases in the midlatitude

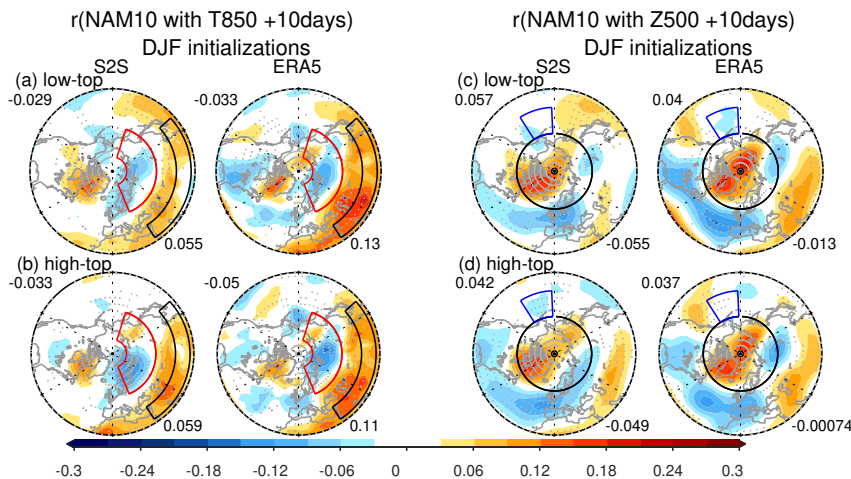


Figure 7. Correlation of NAM10 with T850 and Z500 lagged by 10 days. T850 and Z500 are multiplied by -1 to focus on the anomalies associated with a weak vortex. The low-top multi-model mean (based on 4 models) is shown in the top row, and the high-top multi-model mean (based on 14 models) in the bottom row. Individual models are shown in Supplemental Figure S2 and S4. The correlation is computed for each model separately across all available DJF initializations, and then the correlations are averaged to form multi-model means. For (a) and (b), the area-weighted correlation in the red box over subpolar Eurasia is shown at the top left for each panel, and at the bottom right for the black box over subtropical Eurasia. For (c) and (d), the area-weighted correlation in the polar cap (north of black line) is shown at the top left for each panel, and at the bottom right for the blue box over the North Pacific. Stippling indicates where (left column) all included models agree on the sign of the response, and (right column) all models suffer from the same signed bias.

response to the stratospheric NAM mirror biases in the regional manifestation of the tropospheric NAM in these models. We demonstrate this with the lagged regression of the NAM at 1000 hPa with Z500 and with T850 (Figure 10). While the models capture the signal for Z500 anomalies over the polar cap and if anything represent it too strongly, they underestimate the low in the Azores region and the ridge over southwest Asia. The North Africa and Middle East warming and the subpolar Eurasian cooling, are also underestimated (Figure 10). This similarity between midlatitude biases of the stratospheric NAM vs. tropospheric NAM indicates that the models' inaccuracy in the representation of internal tropospheric dynamics is leading to a poor representation of the regional structure of downward coupling.

An additional bias in the Z500 response to stratospheric variability (Figure 7c,d) is evident in the North Pacific. While the observed response to the polar stratosphere is weak in the Pacific sector, the models simulate a response that resembles the Atlantic sector response, projecting onto the Aleutian Low. The net effect is that the modeled response is too zonal, and insufficiently concentrated in the Atlantic sector. Such a model bias has also been noted in climate models participating in CMIP6 (Dai et al., 2024) and the C3S seasonal forecast database (Lee and Polvani, 2024), and this bias emerges more strongly as lags increase (Figure 8d) and is evident in almost all individual ensemble members (Figure 9d).



Systematic biases in downward coupling

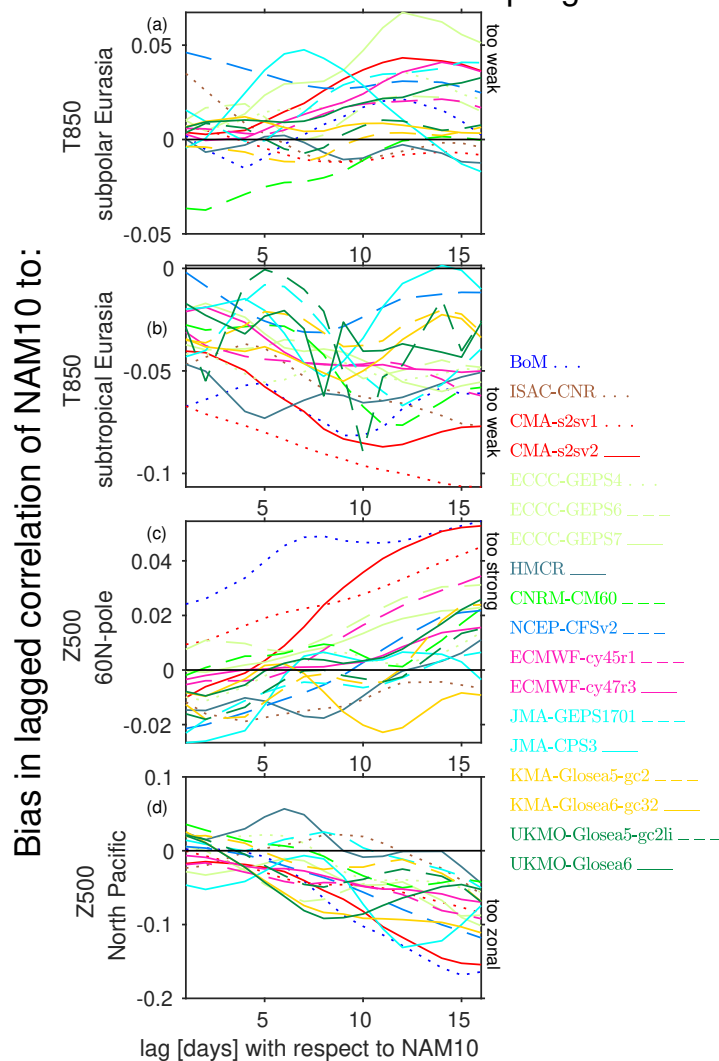


Figure 8. Development of the bias in the ensemble mean of each system (relative to ERA5 subsampled to match each system) of downward coupling of NAM10 [days 9-12] to the tropospheric response lagged up to 16 days later for DJF initializations in the following four regions: (a) temperature at 850 hPa in subpolar Eurasia (see red box on Figure 7ab; 22.5-162.5E, 52.5-77.5N); (b) temperature at 850 hPa in subtropical Eurasia (see black box on Figure 7ab; 32.5-132.5E, 27.5-37.5N); (c) polar cap height at 500 hPa (see black line on Figure 7cd); (d) Z500 in the North Pacific (see blue box on Figure 7cd; 182.5-212.5E, 37.5-57.5N). Low-top models are dotted, and older versions of high-top models are dash-dot.

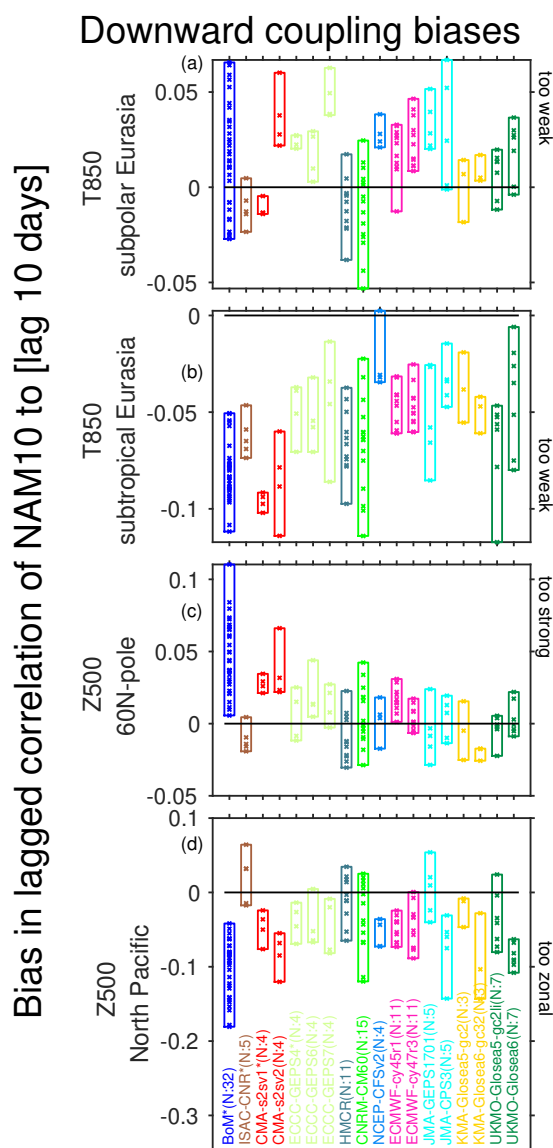


Figure 9. Summary of the bias in each member of each system (relative to ERA5 subsampled to match each system) of downward coupling of NAM10 [days 9–12] to the tropospheric response lagged 10 days later for DJF initializations in the following four regions: (a) temperature at 850 hPa in subpolar Eurasia (see red box on Figure 7ab; 22.5–162.5E, 52.5–77.5N); (b) temperature at 850 hPa in subtropical Eurasia (see black box on Figure 7ab; 32.5–132.5E, 27.5–37.5N); (c) polar cap height at 500 hPa (see black line on Figure 7cd); (d) Z500 in the North Pacific (see blue box on Figure 7cd; 182.5–212.5E, 37.5–57.5N). Low top models are indicated with a star after the name.

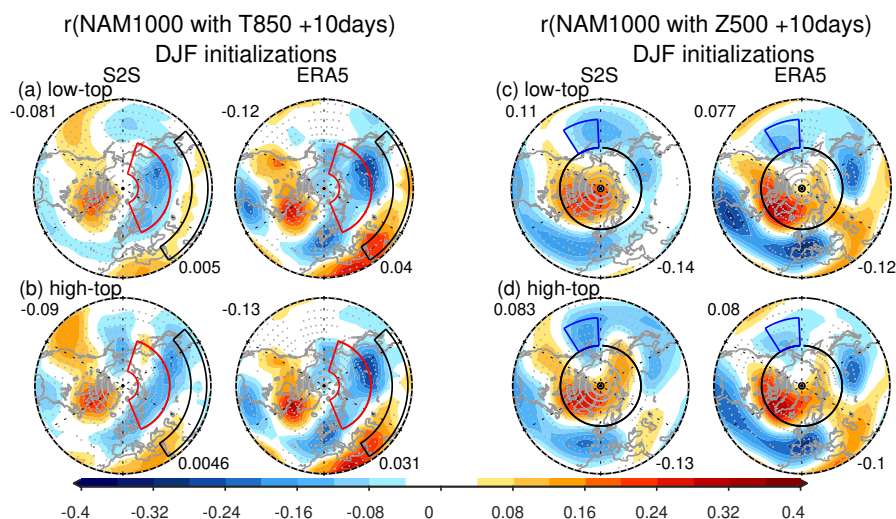


Figure 10. Correlation of NAM1000 with Z500 and T850 lagged by 10 days, with NAM1000 taken in days 9 to 12 after initialization. Other plotting conventions follow Figure 7.

Previous work has noted two possible explanations for this bias. First, Dai et al. (2023) argue that the muted Pacific sector response in observations is linked to North Pacific SSTs associated with the Pacific sector tropospheric precursor that initially drove the vortex anomaly. Namely, a deepened northwest Pacific low helps induce a weaker vortex (Garfinkel et al., 2010; Cho et al., 2022), while the underlying North Pacific SST anomalies act to suppress the Pacific sector response to the vortex anomaly. Building on this, Dai et al. (2024) further show that CMIP6 models underestimate these North Pacific SST anomalies; as a result, they simulate a Pacific sector response to the vortex anomaly that resembles the Atlantic sector response. Second, Lee and Polvani (2024) note that seasonal models suffer from too much variance in the Aleutian low region, and that this region is overly sensitive to zonally symmetric NAM perturbations.

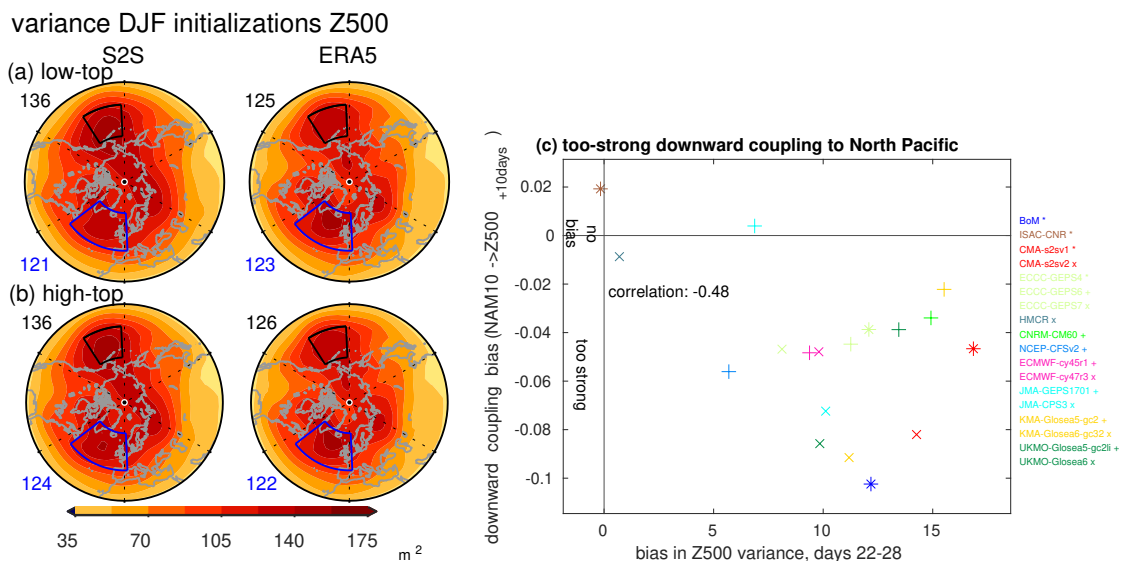


Figure 11. Variance of Z500 in days 22–28 in the low-top (4 models) and hi-top (14 models) multi-model mean and in ERA5 subsampled to match the available systems. The area-weighted average in the box over the North Pacific and over the North Atlantic is shown for each panel. The variance for each model is shown in Supplemental Figure S9. Panel (c) compares the (x-axis) bias in variance of Z500 in the North Pacific box from panels (a-b) to (y-axis) bias in downward coupling of Z10 to Z500 over the North Pacific [182.5–212.5E, 37.5–57.5N; same region as in panels (a-b)] at a lag of 10 days (Figure 7cd). Biases are computed with respect to ERA5 subsampled to match each system.

We find evidence for both of these possibilities. First, we demonstrate in Figure 8d that this bias in downward coupling to the North Pacific is not evident if Z500 is examined simultaneous with NAM10, but rather grows with time as Z500 is lagged relative to NAM10 (Figure 8d; cf. lags of 5 days in Supplemental Figure S6). This implies that models capture the weak response in the Pacific at short lags, and so can capture the initial impact of tropospheric precursors on the immediate Pacific sector response. However, the Pacific sector imprint strengthens at later lags in the subseasonal models, allowing the NAM signal to become increasingly zonal. Note that a similar bias is evident if we examine the lagged correlation of Z500 against NAM1000 (Figure 10): the tropospheric NAM is overly zonal as well at later lags, but at earlier lags this bias goes away (Supplemental Figure S8). There is a significant relationship between this bias in the tropospheric NAM signature in the North Pacific lagged 10 days and the bias in downward coupling ($r = 0.47$): models with too much of a North Pacific low in the negative phase of the tropospheric NAM also simulate overly strong downward coupling.

Next, we consider the effect noted by Lee and Polvani (2024), whereby too-strong variance in the Pacific sector acts to enhance the Pacific center of action of the NAM. Figure 11ab shows the variance in Z500 days 22–28 in DJF for the low-top and high-top multi-model mean and for subsampled ERA5. Each panel also indicates the variance in the North Atlantic and North Pacific regions where variance is strongest. While the models simulate a reasonable variance in the Atlantic, they simulate too much variance in the Pacific sector by $\sim 10\%$ in the multi-model mean, with biases in a few individual models exceeding $\sim 15\%$ (Supplemental Figure S9). Next, we assess whether this bias in Pacific sector variance (x-axis of Figure 11c)

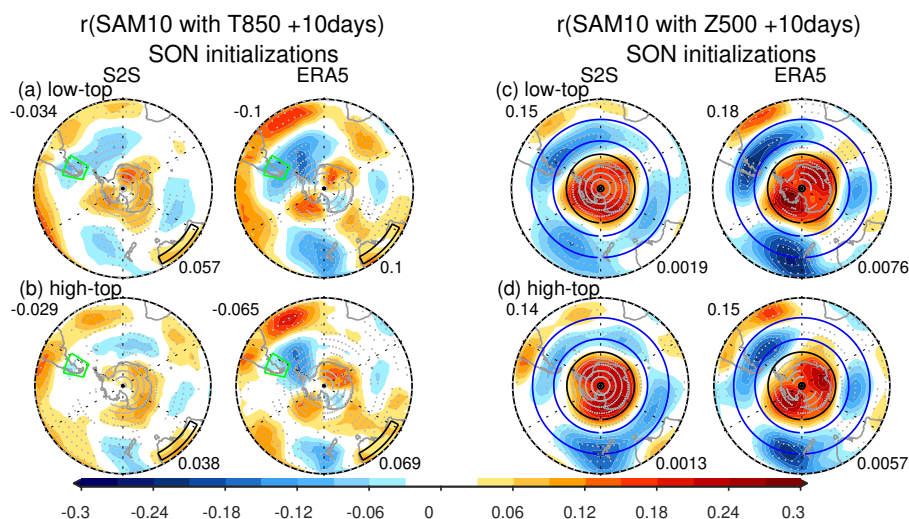


Figure 12. As in Figure 7 but for initializations in SON. The mean correlation over southern Australia (black box) is indicated in the bottom right of panels (a) and (b), and southern South America (green box) is indicated in the top left of panels (a) and (b). The mean correlation for polar cap height (black line) is indicated on the top left of (c) and (d), and the wave-3 amplitude of Z500 in midlatitudes (bounded by blue lines) on the bottom right of (c) and (d). These regions are focused upon in Figures 14 and 13.

is related to biases in Pacific sector downward coupling (y-axis of Figure 11c). The biases are significantly correlated with each other ($r = -0.48$): models with particularly large biases in variance also suffer the most from overly-zonal downward coupling. This effect is generally consistent with Lee and Polvani (2024). Nonetheless, this effect is strongly influenced by the few models with weak bias, and if these were removed, the correlation would change substantially.

5.2.2 Regional structure of the Southern Annular Mode (SAM)

Garfinkel et al. (2025) found that downward coupling over the polar cap in the SH from 10 hPa to 100 hPa was too weak in subseasonal models, while it was too strong from 100 hPa to 850 hPa. In the net, downward coupling from 10 hPa to 850 hPa was reasonably simulated in the multi-model mean. We now consider the regional structure of this downward coupling.

We begin with a map view of the correlation of temperature at 850 hPa anomalies lagged 10 days after the SAM at 10hPa (Figure 12a,b and Supplemental Figure S10). The models simulate the expected temperature impacts associated with the SAM: a negative SAM lead to warm conditions over Australia and midlatitude South America, and cold conditions over the Antarctic Peninsula and southern South America. The magnitude of these impacts in models is too weak, however, in all of these regions in nearly all models (Figure 13a,b) and in nearly all individual ensemble members (Figure 14a,b). Biases are similar in both low-top and high-top models (Figure 12a,b), and also if we consider a regression perspective (Supplemental Figure S11).

These biases are also evident if we evaluate the regional response to the tropospheric SAM: while the warming over Antarctica and cooling over southern South America is evident, they are too weak by a factor of 2 over Australia and four over

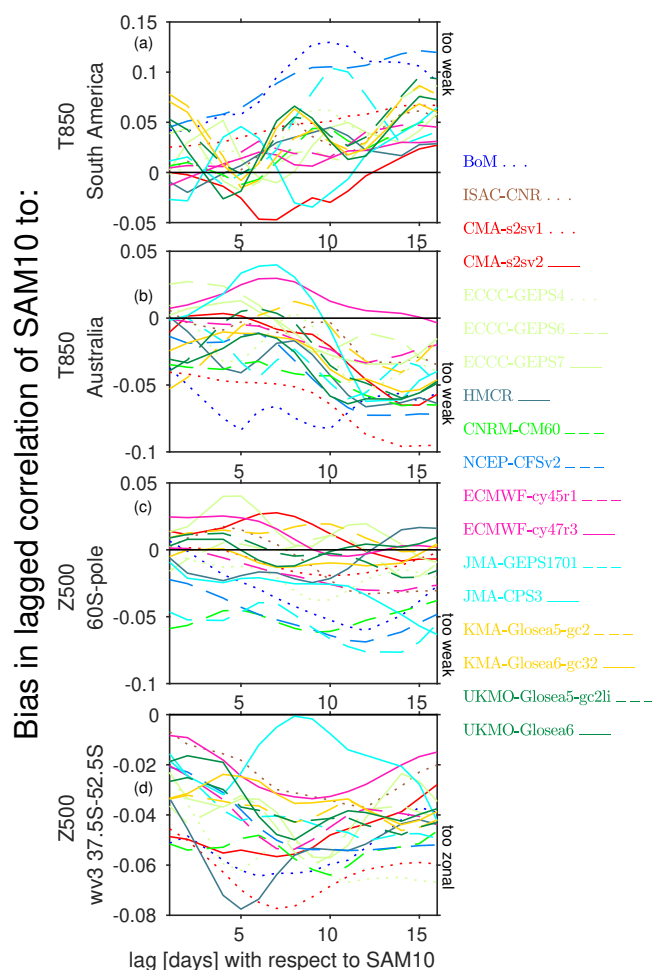


Figure 13. Development of the bias in the ensemble mean of each system (relative to ERA5 subsampled to match each system) of downward coupling of SAM10 [days 9-12] to the tropospheric response lagged up to 16 days later for SON initializations in the following four regions: (a) temperature at 850 hPa in southern South America (see green box on Figure 12ab; 77.5-57.5W, 42.5-57.5S); (b) temperature at 850 hPa in Australia (see black box on Figure 12ab; 117.5-152.5E, 27.5-32.5N); (c) polar cap height at 500 hPa (poleward of black line on Figure 12cd); (d) wave-3 amplitude of Z500 in the SH midlatitudes (37.5-52.5S) (area enclosed by blue lines on Figure 12cd).

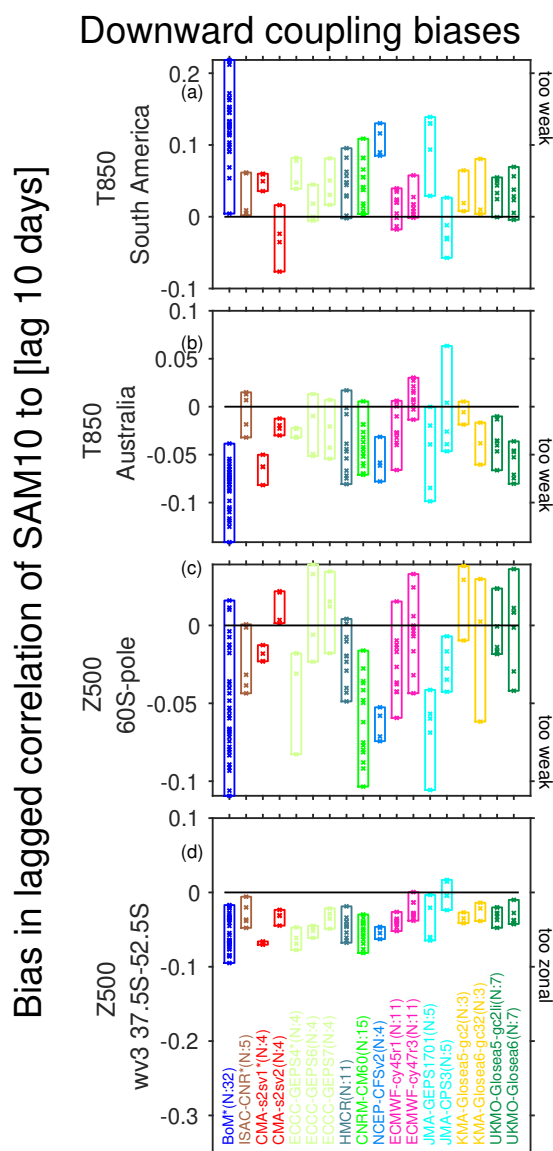


Figure 14. Summary of the bias in each member of each system (relative to ERA5 subsampled to match each system) of downward coupling of SAM10 [days 9–12] to the tropospheric response lagged 10 days later for SON initializations in the following four regions: (a) temperature at 850 hPa in southern South America (see green box on Figure 12ab; 77.5–57.5W, 42.5–57.5S); (b) temperature at 850 hPa in Australia (see black box on Figure 12ab; 117.5–152.5E, 27.5–32.5N); (c) polar cap height at 500 hPa (poleward of black line on Figure 12cd); (d) wave-3 amplitude of Z500 in the SH midlatitudes (37.5–52.5S) (area enclosed by blue lines on Figure 12cd).

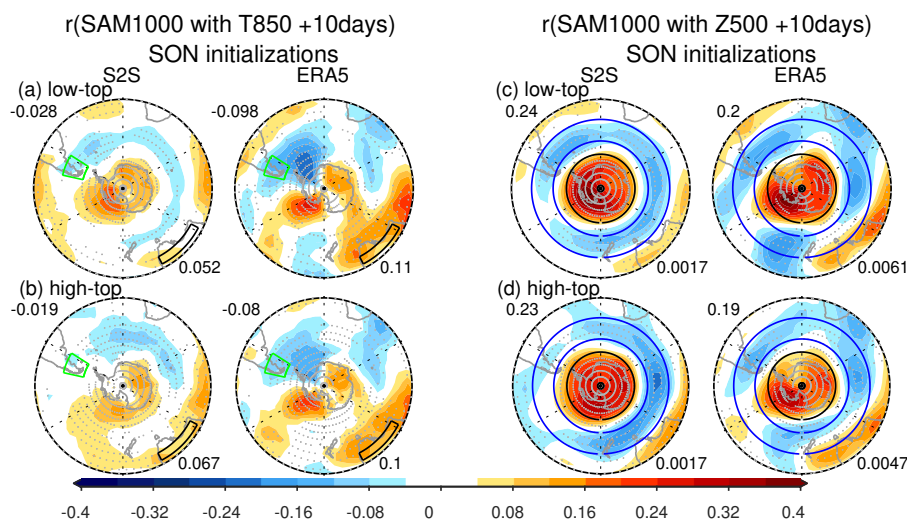


Figure 15. Correlation of SAM1000 with Z500 and T850 lagged by 10 days, with SAM1000 taken in days 9 to 12 after initialization. Other plotting conventions follow Figure 12 .

southern South America (Figure 15ab). The biases are quantitatively similar whether we consider downward propagation from the stratosphere or anomalies associated with the tropospheric SAM. This suggests that the bias is not specific to stratosphere-troposphere coupling per se but rather to an inability to fully represent the regional impacts of the SAM.

A similar story emerges if we analyze Z500 anomalies lagged 10 days after SAM at 10 hPa (Figure 12c,d and Supplemental Figure S12). Over the polar cap, the downward impact is reasonable in the multi-model average, though it is slightly too weak in the majority of models and ensemble members (Figure 13c and 14c). In contrast, the regional wave-3 structure in midlatitudes is too weak in all models and in the multi-model mean, and entirely missing in some models (e.g., ECCC-GEPS6/7, HMCR, CMA-S2Sv1, UKMO-GloSea; Figure 13d and 14d and Supplemental Figure S10). This bias in the midlatitude wave-3 pattern is also evident when considering a regression perspective or at other lags for Z500 (not shown).

The wave-3 pattern is also muted when consider the response to the tropospheric SAM even as the polar cap Z500 signal is well-represented (Figure 15c,d), and so indicates that the difficulty in models is not specific to stratosphere-troposphere coupling but rather is generic to the SAM.

6 Discussion and Conclusions

Operational subseasonal models must capture stratosphere-troposphere coupling processes correctly in order to fully exploit windows of opportunity that arise when the stratospheric polar vortex is weak or strong, as the prediction skill of the tropospheric annular modes is higher (the SAM especially) in most models during these active vortex conditions (Figure 1; Sigmond et al., 2013; Tripathi et al., 2015; Sun et al., 2020; Domeisen et al., 2020b). The previous two papers from the SNAP biases project (Lawrence et al., 2022; Garfinkel et al., 2025) identified several systematic model biases in the mean state of the polar



stratosphere and in coupling processes connecting the tropospheric wave flux, polar cap height, and the stratospheric polar vortex. The goal of this study is to evaluate the connection between these systematic biases across subseasonal forecast systems and forecast skill for the NAM and SAM, and also to consider the regional impacts of biases in downward coupling.

Updated or upgraded model versions from nearly all modeling centers show higher skill in both the stratospheric and tropospheric NAM and SAM as compared to their earlier versions (GloSea a notable exception), and models with relatively higher skill for the stratospheric annular modes generally have higher skill for the tropospheric annular modes as well (Figure 3).

Stratospheric NAM and SAM skill is significantly higher in models that better represent the upward coupling of heat flux from 500 hPa to 100 hPa, and for the NAM the skill is also enhanced in models that better simulate the sensitivity of the polar vortex to 100 hPa heat flux (Figure 4). These correlations are quantitatively similar to the corresponding relationship between NAM skill and biases in climatological wave-1 heat flux at 100 hPa, and between SAM skill and biases in climatological wave-2 heat flux at 100 hPa (Figure 5). In turn, models with better climatological heat flux also simulate better upward coupling (Garfinkel et al., 2025). Hence, it is ambiguous whether stratospheric SAM and NAM skill relies on a better heat flux mean state or upward propagation of wave activity. Either way, these sensitivities are stronger than the non-significant relationship between mean state biases in the vortex and in stratospheric NAM or SAM skill (Figure 5c,i).

Tropospheric NAM and SAM skill do not appear to be directly related to biases in downward coupling or to mean state biases in the lower stratosphere (Figures 4 and 5). However, anomalous states of the polar vortex can open windows of opportunity for more skillful forecasts of the tropospheric NAM on subseasonal timescales (Figure 1). Tropospheric NAM skill is higher in models with higher resolution (Figure 6). Because models with higher vertical resolution in the stratosphere tend to also have higher vertical resolution in the troposphere ($r = 0.67$) it is difficult to isolate the role of resolution in each layer for skill.

Downward coupling over the NH polar cap is too strong in most subseasonal models (Figure 8c, Garfinkel et al., 2025). This bias does not, however, extend to midlatitudes. Rather, downward coupling over Eurasia is too weak in nearly all models, and is particularly poor when considering the surface temperature over subtropical Eurasia in response to the stratospheric polar vortex: it is underestimated by all models by at least 20% and in some models by more than a factor of two (Figure 7, Figure 8a,b). A similar bias is evident when considering the lagged temperature response to tropospheric annular mode variability (Figure 10 and Supplemental Figure S13): the Azores center of action, the southwest Asia Z500 anomaly, and the surface temperature impacts in Eurasia are too weak. This effect is consistent with Garfinkel et al. (2024), who found that anomalous NAM states decay too fast, and that eddy feedback in the North Atlantic is too weak, in a subset of the models considered here.

We find that downward coupling to the North Pacific is too strong in models. This bias is not evident for simultaneous lags, but rather develops as we examine the lagged Z500 response to preceding NAM10 anomalies (Figure 8d). This bias is also evident but weaker when considering the Z500 response to the tropospheric NAM (Figure 10 and Supplemental Figure S14), which suggests that part of the problem might not be in downward coupling per se, but could be a more generic difficulty in representing the zonal structure of the annular modes and unforced variability in this sector (Lee and Polvani, 2024). One possibility is too strong eddy driving in the Pacific (cf., Figure 8a,b of Garfinkel et al., 2024), but analyzing this possibility is beyond the scope of this study. An additional interpretation is that tropical forcing of Rossby waves might be poorly represented by these models, and this missing tropical forcing could weaken the expected meridionally-oriented dipole component centered



in the Pacific and lead to an overestimated zonal component of the flow. For example, although tropical MJO forcing tends to contribute to the predictability of heat flux into the stratosphere at timescales of 2–4 weeks in subseasonal models (Garfinkel and Schwartz, 2017), the ensemble spread of the extratropical impacts at longer lead times hints at large intrinsic forecast uncertainty (Straus et al., 2023) and teleconnections are too weak in nearly all models (Garfinkel et al., 2022; Stan et al., 2022). Future work is needed to consider this possibility.

In the SH, the models realistically represent downward coupling over the polar cap but underestimate the wave-3 pattern in midlatitudes and the temperature impacts over Australia and South America (Figure 12 and 13). This poor representation of midlatitude variability associated with the polar vortex is likely related to a more general inability to represent the midlatitude impacts of the tropospheric SAM since the wave-3 pattern and warming over Australia are also weaker in response to tropospheric SAM variability (Figure 15 and Supplemental Figure S15). These effects are consistent with Garfinkel et al. (2024), who found that eddy feedback in the Southern Hemisphere is too zonally symmetric in a subset of the models considered here.

While models simulate a Eurasian temperature response to stratospheric anomalies that is too weak, this is not necessarily reflected in their representation of extreme events. For example, Erner and Karpechko (2024) showed that ECMWF, UKMO, CNRM, BoM, and JMA tend to predict too many cold spells over subpolar Eurasia during weeks 3–4 following weak polar vortex events, seemingly contradicting the results of the present study. For three of these models (BoM, ECMWF, and UKMO), the same model versions were used in both studies, suggesting that the discrepancy likely arises from differences in methodology. In particular, the models may predict too many cold events without making them too cold, or they may underestimate the warm response to strong polar vortex events, neither of which was tested by Erner and Karpechko (2024). Subsequent evaluations should therefore use a wider range of verification metrics to more comprehensively assess forecast-system properties, such as reliability, that affect forecast quality.

Another aspect to consider in interpreting the results is the possibility of a biased signal-to-noise ratio in the models. Lee and Charlton-Perez (2024) find that the signal-to-noise ratio is biased too high in the polar stratosphere in some models, while the signal-to-noise ratio might be too low in the midlatitude troposphere in others (Garfinkel et al., 2024). This would be in agreement with the fact that downward coupling is too strong over the polar cap – an overly strong polar stratospheric signal gets passed down to the tropospheric polar cap via downward coupling – but not in midlatitudes. A complementary effect that could explain the too-weak and overly zonal impact in midlatitudes is transient eddy feedback: Garfinkel et al. (2024) demonstrated that eddy feedback is too weak in the NH, and too zonal in the SH, in a subset of these models. Further work is needed to also investigate these possibilities. A further effect we have not considered here is the role of planetary waves in mediating the downward vertical coupling to the troposphere. For example, Dunn-Sigouin and Shaw (2015) showed that wave coupling is important during vortex events and that the associated tropospheric response has a strong zonally asymmetric component. More generally, both Boville (1984) and Shaw et al. (2014) argued that planetary-wave coupling provides a mechanism through which stratospheric anomalies and biases can influence the tropospheric circulation, particularly in its zonally asymmetric component. While this is beyond the scope of the present study, it may provide another perspective on the apparent disconnect between the diagnosed zonal-mean downward coupling metrics and tropospheric NAM/SAM predictability, and the zonally asymmetric midlatitude impacts.



While we have identified several systematic biases in the NAM and SAM in subseasonal models, we are using an ensemble
 445 of opportunity and so cannot isolate the role of any specific individual factor. For example, there could be other factors affecting
 the intermodel spread in predictability of the SAM and NAM, besides those considered here (e.g., different hindcast periods
 or teleconnections from the tropics). It is conceivable that skill can be lower in some years than others, and this might alias
 the results. However, the common period across these models is only 6 years (2005–2010), which we consider insufficient to
 robustly evaluate skill. More generally, it is impossible to constrain every possible factor that could affect skill in an ensemble
 450 of opportunity. Only experiments in which the vertical resolution (Xiao et al., 2026) or stationary wave structure are carefully
 controlled can isolate these effects, and thereby confirm the results presented here. Nonetheless, our results are in agreement
 with Xiao et al. (2026), showing that stratospheric variability is more predictable when vertical resolution is enhanced.

7 Outlook

The results presented here suggest at least three important questions which the community could/should focus on:

- 455 1. "What dynamical pathway actually transfers stratospheric predictability to the surface?" Models with higher stratospheric
 NAM/SAM skill generally exhibit higher tropospheric NAM/SAM skill, yet improved zonal-mean downward coupling
 metrics do not appear to translate into improved tropospheric skill.
2. "Why do models systematically produce an overly zonal midlatitude tropospheric Annular Mode response to the strato-
 sphere?" The paper identifies a recurring bias across models, namely a spurious Pacific center of action and underesti-
 460 mated regional impacts over midlatitude Eurasia and in the Southern Hemisphere. Understanding why models struggle to
 represent the regional structure of the annular modes in populated regions may be at least as important as understanding
 the biases in the strength of the coupling itself.
3. "Which model improvements actually matter most for predictability?" Higher skill is associated with better vertical
 resolution, improved mean states, and more realistic coupling diagnostics, but these factors are strongly intertwined.
 465 Determining which model developments are causally responsible for improved predictability remains an important open
 question.

Overall, higher skill in the stratosphere ultimately benefits the troposphere, even though it is not clear what exactly needs
 to be done in a model to achieve this. Nonetheless, it is clear that models still have room to improve their representation of
 the stratosphere, and that doing so will likely lead to an improved skill for surface climate. Isolating this improvement in skill
 470 can be difficult in an ensemble of opportunity, however the SNAPSI project (Hitchcock et al., 2022) also led by SNAP isolates
 this effect and demonstrates its robustness using targeted model experiments (Lee et al., 2025; Ayarzagüena et al., 2026; Feng
 et al., 2025; Seviour et al., 2026).



Data availability. The hindcasts from the S2S database used here are available from the European Center for Medium-range Weather Forecasts, (2026): S2S global sub-seasonal to seasonal ensemble real-time forecasts from multiple NWP centres since 2015. ECMWF Data Store
475 <https://doi.org/10.21957/5ac361bf> (Accessed on 29-06-2026) under the "Reforecasts" S2S set. The NOAA GEFSv12 hindcasts can be obtained from <https://registry.opendata.aws/noaa-gefs-reforecast/>. Hindcasts 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 1. ZDL organized and
480 led the SNAP effort leading to this paper, and also downloaded all of the data. AHB produced the final version of Figure 1. CW produced an earlier version of Figures 4 and 5. All the listed coauthors were active participants in this SNAP community effort and provided comments on the draft manuscript.

Competing interests. D. Domeisen is a member of the editorial board of Weather and Climate Dynamics.

Acknowledgements. This work uses S2S Project data. S2S is a joint initiative of the World Weather Research Programme (WWRP) and the
485 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).

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 Bi-national Science Foundation grant no. 2021714 and National Science Foundation grant no. AGS-2140909). I.E. and A.Y.K are supported
490 by the Research Council of Finland (grants No. 355792 and 373300). The work of M.O. is supported by UBACyT 20020220100075BA, PIP 11220200102038CO and PICT-2021-GRF-TI-00498 projects. The work of A.dlC. is supported by grants PID2022-136316NB-I00 and PID2025-169065NB-I00 funded by MICIU/AEI/10.13039/501100011033 and ERDF/EU. M.A., B.A and N.C. acknowledge the support of grant RETO-PV (PID2024-158151NB-I00) funded by MICIU/AEI/10.13039/501100011033 and ERDF/EU. FMP and JG-S have been partially supported by the Spanish ATLANTE project (PID2019-110234RB-C21) and "Ramón y Cajal" programme (RYC-2016-21181),
495 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. S.-W.S. is funded by the Korea Meteorological Administration Research and Development Program under Grant (RS-2025-02313090). 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. E.D.S is supported by the Research Council of Norway grant 309562 Climate Futures.

500 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.



505 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

- 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.
- Ayarzagüena, B., Butler, A. H., Hitchcock, P., Garfinkel, C. I., Lawrence, Z. D., Ning, W., Rupp, P., Wu, Z., Afargan-Gerstman, H., Calvo, N., de la Cámara, A., Jucker, M., Koren, G., Maeseneire, D. D., Manney, G. L., Osman, M., Taguchi, M., Barton, C., Hong, D.-C., Hyun, Y.-K., Kim, H., Knight, J., Malguzzi, P., Mastrangelo, D., Oh, J., Polichtchouk, I., Richter, J. H., Simpson, I. R., Son, S.-W., Specq, D., and Stockdale, T.: The role of the stratospheric state in upward wave flux prior to Sudden Stratospheric Warmings: a SNAPSI analysis, *Weather and Climate Dynamics*, 7, 411–437, <https://doi.org/10.5194/wcd-7-411-2026>, 2026.
- Baldwin, M. P. and Thompson, D. W. J.: 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.
- 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. H., Charlton-Perez, A., Domeisen, D. I., Garfinkel, C. I., Gerber, E. P., Hitchcock, P., Karpechko, A. Y., Maycock, A. C., Son, S.-W., and Thompson, C.: Chapter 10 - Subseasonal predictability and the stratosphere, in: *Sub-seasonal to Seasonal Prediction (Second Edition)*, edited by Robertson, A. W. and Vitart, F., pp. 355–397, Elsevier, second edition edn., <https://doi.org/https://doi.org/10.1016/B978-0-443-31538-1.00006-3>, 2026.
- Cho, H.-O., Kang, M.-J., Son, S.-W., Hong, D.-C., and Kang, J. M.: A critical role of the North Pacific bomb cyclones in the onset of the 2021 sudden stratospheric warming, *Geophysical Research Letters*, 49, e2022GL099245, 2022.
- Dai, Y., Hitchcock, P., and Simpson, I. R.: Dynamics and impacts of the North Pacific eddy-driven jet response to sudden stratospheric warmings, *Journal of Climate*, 36, 865–884, 2023.
- Dai, Y., Hitchcock, P., and Simpson, I. R.: What Drives the Spread and Bias in the Surface Impact of Sudden Stratospheric Warmings in CMIP6 Models?, *Journal of Climate*, 37, 3917 – 3942, <https://doi.org/10.1175/JCLI-D-23-0622.1>, 2024.
- 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, e2019JD030920, <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., 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, e2019JD030923, <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.
- Edmon Jr, H., Hoskins, B., and McIntyre, M.: Eliassen-Palm cross sections for the troposphere, *Journal of Atmospheric Sciences*, 37, 2600–2616, 1980.



- Erner, I. and Karpechko, A.: Factors influencing subseasonal predictability of northern Eurasian cold spells, *Quarterly Journal of the Royal Meteorological Society*, 150, 2955–2975, <https://doi.org/10.1002/qj.4744>, 2024.
- Feng, K., Rao, J., Garfinkel, C. I., Butler, A. H., Jie, W., Wu, T., Hitchcock, P., Lim, E.-P., Dowdy, A. J., and Seviour, W.: Can stratospheric nudging improve surface predictability? Insights from the 2019 Southern Hemisphere sudden stratospheric warming, *npj Climate and Atmospheric Science*, 8, 353, <https://doi.org/10.1038/s41612-025-01234-2>, 2025.
- Fogt, R. L. and Marshall, G. J.: The Southern Annular Mode: Variability, trends, and climate impacts across the Southern Hemisphere, *WIREs Climate Change*, 11, e652, <https://doi.org/10.1002/wcc.652>, 2020.
- 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. and Schwartz, C.: MJO-related tropical convection anomalies lead to more accurate stratospheric vortex variability in sub-seasonal forecast models, *Geophysical Research Letters*, 44, 10 054–10 062, 2017.
- 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., Chen, W., Li, Y., Schwartz, C., Yadav, P., and Domeisen, D. I. V.: The Winter North Pacific Teleconnection in Response to ENSO and the MJO in Operational Subseasonal Forecasting Models Is Too Weak, *Journal of Climate*, 35, 4413–4430, <https://doi.org/10.1175/JCLI-D-22-0179.1>, 2022.
- 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*, 150, 4417–4436, 2024.
- Garfinkel, C. I., Lawrence, Z. D., Butler, A. H., Dunn-Sigouin, E., Erne, I., Karpechko, A. Y., Koren, G., Abalos, M., Ayarzagüena, B., Barriopedro, D., Calvo, N., de la Cámara, A., Charlton-Perez, A., Cohen, J., Domeisen, D. I. V., García-Serrano, J., Hindley, N. P., Jucker, M., Kim, H., Lee, R. W., Lee, S. H., Osman, M., Palmeiro, F. M., Polichtchouk, I., Rao, J., Richter, J. H., Schwartz, C., Son, S.-W., Taguchi, M., Tyrrell, N. L., Wright, C. J., and Wu, R. W.-Y.: A process-based evaluation of biases in extratropical stratosphere–troposphere coupling in subseasonal forecast systems, *Weather and Climate Dynamics*, 6, 171–195, 2025.
- 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.
- Gerber, E. P., Polvani, L. M., and Ancukiewicz, D.: Annular mode time scales in the Intergovernmental Panel on Climate Change Fourth Assessment Report models, *Geophysical Research Letters*, 35, <https://doi.org/10.1029/2008GL035712>, 2008.
- Gerber, E. P., Baldwin, M. P., Akiyoshi, H., Austin, J., Bekki, S., Braesicke, P., Butchart, N., Chipperfield, M., Dameris, M., Dhomse, S., Frith, S. M., Garcia, R. R., Garny, H., Gettelman, A., Hardiman, S. C., Karpechko, A., Marchand, M., Morgenstern, O., Nielsen, J. E., Pawson, S., Peter, T., Plummer, D. A., Pyle, J. A., Rozanov, E., Scinocca, J. F., Shepherd, T. G., and Smale, D.: Stratosphere-troposphere coupling and annular mode variability in chemistry-climate models, *Journal of Geophysical Research: Atmospheres*, 115, <https://doi.org/10.1029/2009JD013770>, 2010.
- Gillett, N. P., Kell, T. D., and Jones, P.: Regional climate impacts of the Southern Annular Mode, *Geophysical Research Letters*, 33, 2006.
- Guan, H., Zhu, Y., Sinsky, E., Fu, B., Li, W., Zhou, X., Xue, X., Hou, D., Peng, J., Nageswararao, M., et al.: GEFSv12 reforecast dataset for supporting subseasonal and hydrometeorological applications, *Monthly Weather Review*, 150, 647–665, 2022.



- Hamill, T. M., Whitaker, J. S., Shlyueva, 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*, 150, 59–79, <https://doi.org/10.1175/MWR-D-21-0023.1>, 2022.
- 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., 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.
- Hitchcock, P., Butler, A., Charlton-Perez, A., Garfinkel, C. I., Stockdale, T., Anstey, J., Mitchell, D., Domeisen, D. I. V., Wu, T., Lu, Y., Mastrangelo, D., Malguzzi, P., Lin, H., Muncaster, R., Merryfield, B., Sigmond, M., Xiang, B., Jia, L., Hyun, Y.-K., Oh, J., Specq, D., Simpson, I. R., Richter, J. H., Barton, C., Knight, J., Lim, E.-P., and Hendon, H.: Stratospheric Nudging And Predictable Surface Impacts (SNAPSI): a protocol for investigating the role of stratospheric polar vortex disturbances in subseasonal to seasonal forecasts, *Geoscientific Model Development*, 15, 5073–5092, <https://doi.org/10.5194/gmd-15-5073-2022>, 2022.
- Hong, D.-C. and Son, S.-W.: Surface Amplification of Stratospheric Sudden Warming by Poleward Mass Flux in the Upper Troposphere and Lower Stratosphere, *Journal of Climate*, 38, 5069–5082, <https://doi.org/10.1175/JCLI-D-24-0597.1>, 2025.
- Karpechko, A. Y., Butler, A. H., and Vitart, F.: Signal, noise and skill in sub-seasonal forecasts: the role of teleconnections, *Weather and Climate Dynamics*, 6, 1661–1681, <https://doi.org/10.5194/wcd-6-1661-2025>, 2025.
- 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*, 3, 977–1001, <https://doi.org/10.5194/wcd-3-977-2022>, 2022.
- Lee, R. W. and Charlton-Perez, A. J.: Diversity of stratospheric error growth across subseasonal prediction systems, *Geophysical Research Letters*, 51, e2023GL107 574, 2024.
- Lee, R. W., Charlton-Perez, A. J., and Lee, S. H.: Stratospheric Impacts on Weather Regimes Following the 2018 and 2019 Sudden Stratospheric Warmings, *Geophysical Research Letters*, 52, e2025GL115 668, <https://doi.org/10.1029/2025GL115668>, 2025.
- Lee, S. H. and Polvani, L. M.: Large model biases in the Pacific centre of the Northern Annular Mode due to exaggerated variability of the Aleutian Low, *Quarterly Journal of the Royal Meteorological Society*, 150, 4478–4497, 2024.
- 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.
- Purich, A., Arblaster, J. M., Boschat, G., Gillett, Z. E., Hobbs, W., Jucker, M., Lim, E.-P., Udy, D., Abram, N., Campitelli, E., et al.: Southern Annular Mode dynamics, projections and impacts in a changing climate, *Nature Reviews Earth & Environment*, 7, 24–42, <https://doi.org/10.1038/s43017-025-00746-y>, 2026.
- 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 Forecasting*, 37, 797 – 815, <https://doi.org/10.1175/WAF-D-21-0163.1>, 2022.
- 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.
- Seviour, W. J. M., Finkel, J., Rupp, P., Mudhar, R., Butler, A. H., Garfinkel, C. I., Hitchcock, P., Ayarzagüena, B., Hong, D.-C., Hyun, Y.-K., Kim, H., Lim, E.-P., Maeseneire, D. D., Messori, G., Koren, G., Sigmond, M., Simpson, I. R., and Son, S.-W.: Forecast-based attribution



- of the role of stratospheric variability in weather extremes, *Weather and Climate Dynamics*, <https://doi.org/10.5194/egusphere-2026-230>,
 620 accepted, 2026.
- Shaw, T. A., Perlwitz, J., and Weiner, O.: Troposphere-Stratosphere Coupling: Links to North Atlantic Weather and Climate, Including Their Representation in CMIP5 Models, *Journal of Geophysical Research: Atmospheres*, 119, 5864–5880, <https://doi.org/10.1002/2013JD021191>, 2014.
- Sigmond, M., Scinocca, J. F., Kharin, V. V., and Shepherd, T. G.: Enhanced Seasonal Forecast Skill Following Stratospheric Sudden Warmings, *Nature Geosci*, 6, 98–102, <https://doi.org/10.1038/ngeo1698>, 2013.
 625
- Simpson, I. R., Hitchcock, P., Shepherd, T. G., and Scinocca, J. F.: Stratospheric variability and tropospheric annular-mode timescales, *Geophysical Research Letters*, 38, <https://doi.org/10.1029/2011GL049304>, 2011.
- Son, S.-W., Kim, H., Song, K., Kim, S.-W., Martineau, P., Hyun, Y.-K., and Kim, Y.: Extratropical Prediction Skill of the Subseasonal-to-Seasonal (S2S) Prediction Models, *Journal of Geophysical Research: Atmospheres*, 125, e2019JD031273, <https://doi.org/10.1029/2019JD031273>, 2020.
 630
- 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.
- Straus, D. M., Domeisen, D. I., Lock, S.-J., Molteni, F., and Yadav, P.: Intrinsic predictability limits arising from Indian Ocean Madden-Julian oscillation (MJO) heating: effects on tropical and extratropical teleconnections, *Weather and Climate Dynamics*, 4, 1001–1018, 2023.
 635
- Sun, L., Perlwitz, J., Richter, J. H., Hoerling, M. P., and Hurrell, J. W.: Attribution of NAO Predictive Skill Beyond 2 Weeks in Boreal Winter, *Geophysical Research Letters*, 47, e2020GL090451, <https://doi.org/10.1029/2020GL090451>, 2020.
- Thompson, D. W. and Wallace, J. M.: The Arctic Oscillation signature in the wintertime geopotential height and temperature fields, *Geophysical research letters*, 25, 1297–1300, 1998.
 640
- Thompson, D. W. and Wallace, J. M.: Annular modes in the extratropical circulation. Part I: Month-to-month variability, *Journal of climate*, 13, 1000–1016, 2000.
- Thompson, D. W., Lee, S., and Baldwin, M. P.: Atmospheric processes governing the northern hemisphere annular mode/North Atlantic oscillation, in: *The North Atlantic Oscillation: Climatic Significance and Environmental Impact*, edited by Ottersen, G., Hurrell, J. W., Visbeck, M., and Kushnir, Y., vol. 134, pp. 81–112, Wiley Online Library, <https://doi.org/10.1029/134GM05>, 2003.
 645
- Thompson, D. W. J. and Wallace, J. M.: Regional climate impacts of the Northern Hemisphere annular mode, *Science*, 293, 85–89, <https://doi.org/10.1126/science.1058958>, 2001.
- Tripathi, O. P., Charlton-Perez, A., Sigmond, M., and Vitart, F.: Enhanced Long-Range Forecast Skill in Boreal Winter Following Stratospheric Strong Vortex Conditions, *Environ. Res. Lett.*, 10, 104007, <https://doi.org/10.1088/1748-9326/10/10/104007>, 2015.
- 650 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>,
 655 2017.



- Wicker, W., Polichtchouk, I., and Domeisen, D. I. V.: Increased Vertical Resolution in the Stratosphere Reveals Role of Gravity Waves after Sudden Stratospheric Warmings, *Weather and Climate Dynamics*, 4, 81–93, <https://doi.org/10.5194/wcd-4-81-2023>, 2023.
- Wu, Z. and Reichler, T.: Variations in the Frequency of Stratospheric Sudden Warmings in CMIP5 and CMIP6 and Possible Causes, *Journal of Climate*, 33, 10 305 – 10 320, <https://doi.org/10.1175/JCLI-D-20-0104.1>, 2020.
- 660 Xiao, H., Wang, L., Chen, X., Scaife, A. A., Hardiman, S. C., Dunstone, N. J., and Liang, Y.: Finer Vertical Resolution Improves the Sudden Stratospheric Warming Prediction Through Better Representing Planetary Waves, *Geophysical Research Letters*, 53, e2025GL120 534, <https://doi.org/10.1029/2025GL120534>, 2026.