An evaluation of the large scale atmospheric circulation and its variability in the Community Earth System Model 2 (CESM2) and other CMIP models

Isla R. Simpson[1], Julio Bacmeister[1], Richard B. Neale[1], Cecile Hannay [1], Andrew Gettelman[1], Rolando, R. Garcia[2], Peter H. Lauritzen[1], Daniel, R. Marsh[2], Michael J. Mills[2], Brian Medeiros[1], Jadwiga H. Richter[1]

[1]Climate and Global Dynamics Laboratory, National Center for Atmospheric Research, Boulder, CO
[2]Atmospheric Chemistry Observations and Modelling Laboratory, National Center for Atmospheric Research, Boulder, CO

Key Points:

- CESM2 is compared to CESM1 and other CMIP5 and CMIP6 models and ranks highly in many regards.
- Jets, stormtracks, stationary waves, divergent circulation, NAM, SAM, NAO and blocking are assessed.
- CESM2 has improvements in stormtracks and NH winter circulation but degradations in SH circulation.

Corresponding author: Isla Simpson, `islas@ucar.edu`

Abstract

The Community Earth System Model 2 (CESM2) is the latest Earth System Model developed by the National Center for Atmospheric Research in collaboration with the university community and is significantly advanced in most components compared to its predecessor (CESM1). Here, CESM2s representation of the large scale atmospheric circulation and its variability is assessed. Further context is provided through comparison to the CESM1 large ensemble and other models from the Coupled Model Intercomparison Project (CMIP5 and CMIP6). This includes an assessment of the representation of jet streams and storm tracks, stationary waves, the global divergent circulation, the annular modes, the North Atlantic Oscillation and blocking. Compared to CESM1, CESM2 is substantially improved in the representation of the storm tracks, Northern Hemisphere (NH) stationary waves, NH winter blocking and the global divergent circulation. It ranks within the top 10% of CMIP-class models in many of these features. Some features of the Southern Hemisphere (SH) circulation have degraded, such as the SH jet strength, stationary waves and blocking, although the SH jet stream is placed at approximately the correct location. This analysis also highlights systematic deficiencies in these features across the new CMIP6 archive, such as the continued tendency for the SH jet stream to be placed too far equatorward, the North Atlantic westerlies to be too strong over Europe, the storm tracks as measured by low level meridional wind variance to be too weak and a lack of blocking in the North Atlantic sector.

1 Introduction

The Community Earth System Model, version 2 (CESM2) is the second generation Earth System Model developed by the United States' National Center for Atmospheric Research (NCAR), in collaboration with university researchers (Hurrell et al., 2013). Prior to the first incarnation of CESM (CESM1), the history of development of this model can be traced through the Community Climate System Model, versions 4 (CCSM4, Gent et al., 2011), 3 (CCSM3, Collins et al., 2006), 2 (CCSM2 Kiehl & Gent, 2004), the Climate System Model 1 (CSM1 Boville & Gent, 1998) and before that, the Community Climate Model, versions 3 (CCM3 Kiehl et al., 1998), 2 (CCM2 Hack et al., 1993), 1 (CCM1 Williamson et al., 1987) and 0 (CCM0 Washington, 1982; Williamson, 1983). As such, CESM2 represents the current state-of-the-art in Earth System Modelling from this center, incorporating model development contributions from over four decades of research and the efforts of countless individuals.

Over this development history, the array of complex atmospheric, oceanic, hydrologic, cryospheric and biogeophysical processes represented by this model has made CESM2 one of the most comprehensive and complex Earth System Models (ESMs) available. Given its fundamental role in the Earth System, the large scale atmospheric circulation has been represented with some realism, relatively speaking, since the earliest days of climate modeling. Nevertheless, persistent biases remain in certain aspects and, as our models increase in complexity, we must continue to strive for the greatest accuracy possible in the representation of this underpinning feature of the Earth System.

In this study we present an evaluation of basic features of the large scale atmospheric circulation and its variability in CESM2. We provide context by assessing changes compared to its predecessor (CESM1) and by placing it within the wider distribution of Earth System Models as represented by those participating in the Coupled Model Intercomparison Project, phases 5 and 6 (CMIP5 (Taylor et al., 2012) and CMIP6 (Eyring et al., 2016)). The range of atmospheric circulation features presented here is not exhaustive and the primary focus is on the global climatology of the divergent circulation and stationary waves, mid-latitude jet streams and storm tracks and aspects of extra-tropical variability. Separate studies in this special issue provide an assessment of tropical intrasea-

sonal variability, monsoons (Meehl et al., 2020) and El Niño-Southern Oscillation (ENSO) variability and its teleconnections (Capotondi et al., 2020).

Rather than taking the traditional approach of providing an overall introduction, methodological description and summary of the results, we instead provide a self contained introduction and methodology within each results section for each feature considered, such that a reader can easily find all the relevant information in one place for their feature of interest. This diagnostic analysis is intended primarily as a resource for CESM2 users but also serves as a concise summary of the representation of these features in CMIP5 and CMIP6 models.

We begin by describing the model simulations and observational datasets in section 2, followed by a description of the error metrics used and the uncertainty assessments performed in section 3. In section 4 we discuss the representation of jet streams and storm tracks, in section 5 we discuss stationary waves and the global divergent circulation and in section 6 we assess the annular modes, North Atlantic Oscillation (NAO) and blocking. Summary and conclusions are provided in section 7.

2 Model simulations and observation based datasets

For each of the model historical simulations and reanalyses described below, our primary focus is on the period from 1979 to 2014 and on monthly and daily averaged fields of zonal wind (ua), meridional wind (va), geopotential height (zg) and sea level pressure (slp). Note that here we are using variable names as specified by CMIP as opposed those used in CESM2. Each of these fields is first regridded to a common 2 degree horizontal grid using bilinear interpolation before any other fields or metrics are derived. Only the summer and winter seasons are considered in the main text, but equivalent figures are shown for the spring and autumn in the supplementary material.

2.1 CESM2

In its default configuration, CESM2 simulates the global coupled Earth System at approximately $1°$ horizontal resolution. It contains interactive components for the atmosphere, land, ocean, sea ice, river transport and land ice. CESM2 represents a significant advance over CESM1 in many ways (see Danabasoglu et al. (2019) for more details). As the updates within the atmosphere component (Community Atmosphere Model 6, CAM6) are likely to be the most relevant, we summarize some of those major changes here, but readers are referred to Bogenschutz et al. (2018) and Gettelman et al. (2019) for a more detailed description of CAM6 and the high-top atmospheric component (Whole Atmosphere Community Climate Model, WACCM6), respectively.

In the transition from CAM5 to CAM6, almost every physical parameterization within the atmosphere has been updated, with the exception of radiation. A major change is that the boundary layer, shallow convection and cloud macrophysics are combined within the new Cloud Layers Unified By Binormals (CLUBB) scheme (Golaz et al., 2002), resulting in a more consistent representation of boundary layer turbulence (Bogenschutz et al., 2013). The prognostic cloud microphysics scheme (MG2, Gettelman & Morrison, 2015) has been updated from its predecessor (MG1) with a major change being the inclusion of prognostic precipitation. Finally, and of relevance to some of the following results, there have been major updates to the representation of orographic drag. The orographic gravity wave drag scheme now includes a representation of the orientation of sub-grid orography (ridges) and the effects of mesoscale orographic blocking. Furthermore, the turbulent orographic form drag scheme has been updated from the Turbulent Mountain Stress (TMS, Richter et al., 2010) parameterization to that of Beljaars et al. (2004).

Our primary focus will be on four CESM2 historical ensembles that differ in the vertical extent of the atmospheric component and in the presence or absence of coupling to the fully dynamic ocean model. These ensembles are summarized in the lower left of Table 1 and a more detailed description is provided in Table 2. Eleven members make up the ensemble BCAM6 in which the low-top atmosphere model (CAM6), with 32 layers in the vertical extending to ∼40km, is coupled to the ocean model. A three member coupled ensemble with the high-top WACCM6 (Gettelman et al., 2019), which has 70 levels in the vertical extending to ∼130km, will be referred to as BWACCM6. Spatial maps in the main text are only shown for BCAM6, but equivalent figures for BWACCM6 are shown in supplemental Figs S16 and S17. In addition, there are three member ensembles, with prescribed historical sea surface temperatures (SSTs) (Hurrell et al., 2008), referred to as FCAM6 and FWACCM6, for CAM6 and WACCM6 respectively. In this naming convention, B refers to the CESM B-component set which includes coupling to the ocean model, while F refers to the CESM F-component set where SSTs and sea ice are prescribed (i.e., AMIP-type simulations).

These simulations are run under historical forcings (van Marle et al., 2017; Hoesly et al., 2018) until 2014. The coupled simulations are each initialized from different years from a pre-industrial (i.e., perpetual year 1850 forcing) control that has been spun-up for over 1000 years (Danabasoglu et al., 2019), while the prescribed SST simulations begin in 1950. For each ensemble we will only consider the period from 1979 to 2014 for comparison with modern reanalyses over the satellite era.

In addition to these four ensembles of simulations which are contributed to the CMIP6 archive, we will make use of the following simulations that are designed to isolate the underlying cause of some of the changes found in CESM2. FCAM6MOD is an historical simulation with CAM6 with prescribed SSTs but with the SSTs taken from one of the coupled BCAM6 members, as opposed to observations. We will use 1979-2014 of this simulation to explore the role of SST differences versus the lack of coupling in explaining differences between BCAM6 and FCAM6. To explore the influence of changes in orographic parameterizations schemes, four single member experiments, under historical forcings from 1979-2005, with SSTs prescribed to observations will be considered. FCAM6* is an uncoupled simulation with prescribed observed SSTs, very similar to FCAM6 described above, but with biogeochemistry in the land turned off. While the issue of land biogeochemistry is not important for our purposes, we use this rather than FCAM6 for like-with-like comparison with each of the following experiments that also have biogeochemistry turned off. FCAM5 is a simulation performed in the same way as FCAM6* (same forcings, boundary conditions and land model) but with CAM5 physics used instead of CAM6. This allows for an assessment of the overall influence of the atmospheric physics package in isolation, which can then be compared with the following two experiments to isolate the orographic influence. FCAM6_TMS is as FCAM6* but with the new Turbulent Orographic Form Drag (TOFD) scheme of (Beljaars et al., 2004) replaced by the older Turbulent Mountain Stress (TMS) parameterization of CAM5. A comparison of FCAM6* with FCAM6_TMS demonstrates the influence of this change in TOFD. FCAM6_NOMOB is as FCAM6* but without the new Mesoscale Orographic Blocking parameterization included. A comparison between FCAM6* and FCAM6_NOMOB indicates the influence of the new mesoscale orographic blocking scheme.

2.2 CESM1

To examine the changes that have arisen as a result of the developments in advancing from CESM1 to CESM2 and to provide an indication of the sampling uncertainty in each metric as a result of internal variability, we will compare with the CESM1 large ensemble (Kay et al., 2014). This 40 member ensemble of simulations is initialized in 1920 from a single state, with ensemble spread introduced through a round-off level noise perturbation added to the temperature field at initialization. The initial state is that of a

single realization that was branched from an 1850's control simulation and run until 1920 under historical forcings (Lamarque et al., 2010). The 40 member ensemble is then run under historical forcings to 2005 and RCP8.5 forcings, thereafter (Meinshausen et al., 2011; Lamarque et al., 2011). We will assess the 1979-2014 period using the historical and RCP8.5 simulations combined and this will be referred to as LENS.

2.3 CMIP5

As with LENS, we will combine years 1979-2005 of the historical simulations with years 2006-2014 of the RCP8.5 simulations for the 35 CMIP5 models listed in Table 1. For monthly data we make use of all available ensemble members that have both historical and RCP8.5 components, resulting in ensemble sizes ranging from 1 to 10 members (third column of Table 1). We will always show the ensemble mean of a metric for each model when multiple members are available. Error metrics are first calculated for individual members before the ensemble averaging is performed so as to avoid comparing smoother ensemble mean spatial fields with the noisier fields of individual members. For metrics that involve daily ua and va data, we use one member from the 16 models (highlighted with a * in Table 1) and for daily zg data we use one member from the 15 models (highlighted with a + in Table 1). For models that have more than one member available, we use the member with the lowest realization number. Daily fields are obtained by averaging 6 hourly pressure level fields. In each figure, a CMIP5 model can be identified by the model number given in the left column of Table 1.

2.4 CMIP6

For CMIP6, we make use of 1979-2014 of the historical simulations, run under the same forcings as the CESM2 simulations described above. At the time of writing, 42 models are available with ensemble sizes ranging from 1 to 32 (Table 1, right three columns). While the BCAM6 and BWACCM6 ensembles are contributed to the CMIP6 archive, we consider them separately here. Only a subset of 27/20 models, highlighted with a */+ in Table 1, have daily averaged $(ua,va)/zg$ data available and for each of these we only use one member (the member with the lowest realization number).

In each figure, a CMIP6 model can be identified by the model number given in the third from right column of Table 1. We only show error metric summaries for the CMIP6 models in the main text, but ensemble mean spatial bias maps along with indications of model consensus are provided in the Appendix.

2.5 Observation based datasets

Our primary observational comparison will be with atmospheric reanalyses. The new ERA5 reanalysis (C3S, 2019) will be taken as the observational baseline and all simulations and other reanalysis products will be compared to that. Three other modern reanalyses that assimilate a wide array of observations will also be shown: ERA-Interim (Dee et al., 2011); MERRA2 (Gelaro et al., 2017); and JRA55 (Kobayashi et al., 2015). Two twentieth century reanalysis products: ECMWF's ERA20C (Poli et al., 2016) and NOAA's twentieth century reanalysis, 20CR (Compo et al., 2011) are considered, partly for the purpose of assessing those reanalysis products compared to others, but also for the purpose of providing a longer term context for the observational record in certain metrics. These twentieth century reanalyses are only constrained by surface pressure observations (and marine surface winds in the case of ERA20C) and, therefore, lack the additional constraint arising from the multitude of other observations that are assimilated in the other products.

For the most part, we only use 1979 to 2014 for these reanalysis products for direct comparison with the model simulations. MERRA2 only starts in 1980, so for that

product we use 1980 to 2014 and ERA20C only extends to 2010, so for that product we use 1979 to 2010. For the North Atlantic, where a relatively large number of surface pressure observations constrain the 20th century reanalyses back to 1900, we provide an assessment in the variability of metrics using all overlapping 36 year segments between 1900 and 2014 (2010 for ERA20C). For metrics involving daily data, we do not make use of the 20th century reanalyses.

3 Error metric and uncertainty assessments

3.1 Normalized Mean Square Error Metric

When assessing the error in a spatial field (X) we will use the Normalized Mean Square Error ($NMSE$) metric proposed by Williamson (1995). This metric has been applied to the geopotential height field in evaluations of previous NCAR models (Kiehl et al., 1998; Collins et al., 2006; Neale et al., 2013) but here we apply it to all spatial fields considered. The $NMSE$ of the model field X_m is given by

$$NMSE(X_m) = \frac{\overline{(X_m - X_o)^2}}{\overline{(X_o')^2}},\tag{1}$$

where X_o refers to the "observed" field (in our case ERA5), the overbar refers to the area weighted spatial average and the prime refers to the deviation therefrom. To give some indication of where the errors are coming from, the $NMSE$ can be further decomposed into three components, as follows

$$NMSE(X_m) = U + C + P\tag{2}$$

with

$$U = \left(\frac{\overline{X_m} - \overline{X_o}}{\sigma_o}\right)^2 \qquad C = \left(r_{mo} - \frac{\sigma_m}{\sigma_o}\right)^2 \qquad P = \left(1 - r_{mo}^2\right),\tag{3}$$

where σ_o and σ_m refer to the spatial standard deviation of the observations and model, respectively, and r_{mo} refers to the spatial correlation between the model and observations. For a derivation of this, see Murphy (1988), their equation 10, which is essentially the same as this but without normalization of the Mean Squared Error. The first term, U, is the unconditional bias which is a non-dimensional measure of the overall bias in the spatial mean of a field. The second term, C, is the conditional bias, which is non-dimensional and arises through both amplitude and phase errors. It is only non-zero if the regression of X_m onto X_o yields a slope of 1 i.e., if X_m is perfectly correlated with X_o and their spatial variances are equal. The third term is the phase error P, which arises only from errors in the phasing of the spatial variations.

If $r_{mo}=1$ i.e., the phase error is zero, the interpretation of any conditional bias is straightforward; it arises if the amplitude of the spatial variations are too large or too small. When the phase error is non-zero, the interpretation of the conditional bias is less straightforward as it arises through both amplitude and phase errors. Furthermore, the conditional bias can be artificially reduced through a lack of spatial variance. Interpretation can, therefore, be aided by consideration of the scaled variance ratio (SVR) given by

$$SVR(X_m) = \left(\frac{\sigma_m}{\sigma_o}\right)^2 NMSE(X_m)\tag{4}$$

which indicates whether conditional bias arises from too much ($SVR > NMSE$) or too little ($SVR < NMSE$) spatial variance. When $SVR < NMSE$, it also provides caution that the conditional bias component may be artificially reduced through the lack of spatial variance.

Altogether, U, C, P and SVR provide an indication of the roles of an overall mean bias, biases that arise due to errors in the amplitude of spatial variations and biases that

arise due to phasing errors. Fig. 1 provides an explanatory key for how this will be represented in each figure. The *NMSE* will be depicted in each plot with a vertical bar composed of three different colored components for U, C and P while the SVR will be depicted by a circular symbol. When the SVR symbol lies above/below the bar the SVR is greater/less than 1. For models with a spatial mean (unconditional) bias with a magnitude greater than 10% of the ERA5 spatial mean, we depict whether that bias is positive or negative by shading the SVR symbol red or blue respectively.

3.2 Assessment of uncertainty due to internal variability

The 36 year observational record that the models are being compared to, will be subject to uncertainty due to the sampling of internal variability. To provide some indication of the magnitude of this effect, or the significance of differences between the model and the reanalysis, we take a number of approaches:

- When assessing the bias of BCAM6 or LENS relative to ERA5 in map form, we provide an assessment of whether ERA5 lies within the distribution of the 11(40) ensemble members for BCAM6(LENS) and assume that where this is not the case, there is a significant difference between the real world and the model. This is equivalent to a significant difference at the $\sim 9\%(2.5\%)$ level for BCAM6(LENS) by a one-sided non parametric test. Regions where ERA5 does not lie outside of the model ensemble spread will be shaded gray in each figure.
- For each metric we show the minimum to maximum range of that metric from the 40 LENS members, giving an indication of the range of values that can arise due to internal variability within CESM1.
- For each metric we show each of the individual BCAM6, BWACCM6, FCAM6 and FWACCM6 members, giving an indication of the range of values that can arise due to internal variability for CESM2, although this is limited by the relatively small ensemble sizes for these simulations.
- For the Northern Hemisphere (NH) jet stream metrics and the NAO, where there is confidence in the reanalysis extending back to 1900 (Simpson et al., 2019), we show the range of metrics for overlapping observed 36 year climatologies extending back to 1900.

4 Jet Streams and Storm Tracks

4.1 Southern Hemisphere

The Southern Hemisphere (SH) mid-latitude jet stream is important for the representation of weather and surface climate in the SH but also has global implications through the wind stress influence on southern ocean upwelling (Marshall & Speer, 2012; Gent, 2016) and the leakage of warm salty water from the Indian Ocean into the Atlantic through the Agulhas current (Biastoch et al., 2009; Tim et al., 2019). Yet, the SH jet has proven notoriously difficult to model with accuracy, with the majority of models in previous generations positioning the jet stream too far equatorward (Fyfe & Saenko, 2006; Russell et al., 2006; Wilcox et al., 2012; Bracegirdle et al., 2013). Less consistency has been found in model biases in SH jet strength (Russell et al., 2006), although many models exhibit a weaker jet stream and wind stress than observed in the Pacific sector (Fyfe & Saenko, 2006; Bracegirdle et al., 2013). Aside from its implications for the representation of present-day climate, the modelled position of the SH westerly jet stream correlates with its projected poleward shift in response to both ozone depletion (Son et al., 2010) and greenhouse gas forcing (Kidston & Gerber, 2010), primarily during the winter season (Simpson & Polvani, 2016). Given that models with lower latitude jets that sit further from the observed location tend to show a larger poleward shift under forcing, this implies that many models may be over-predicting such responses, although the dynamics behind this remain to be understood (Simpson & Polvani, 2016).

Intrinsically linked to the SH jet stream, is the SH storm track, composed of transient synoptic scale baroclinic eddies that are responsible for the high and low pressure systems that bring day-to-day weather variability to the mid-latitudes and are an important contributor to hemispheric energy, momentum and moisture transports (Chang et al., 2002; Shaw et al., 2016). An equatorward bias in the SH storm track accompanied that of the SH jet stream in both CMIP3 (Chang et al., 2013) and CMIP5 (Chang et al., 2012) models. In terms of storm track strength, the modelled representation has been found to be rather varied in previous intercomparisons but with a preponderance for the SH storm track to be too weak (Chang et al., 2012, 2013).

Inaccuracies in the simulation of the SH jet stream and storm tracks can have many origins: errors in the representation of clouds and radiation can impact on the equator-to-pole temperature gradient and hence baroclinicity (Ceppi et al., 2012); orographic processes have been shown to affect jet latitude (Pithan et al., 2016); and increasing resolution is often accompanied by an increase in the stormtrack strength (Chang et al., 2013; Hertwig et al., 2015). Meehl et al. (2019) recently assessed the influence of resolution and model physics in a series of CESM1 configurations and found comparable effects on storm track intensity arising from changes in both resolution and the representation of cloud radiative effects.

Characteristics of the tropospheric zonal mean mid-latitude westerlies can first be assessed from Fig. 2 (see Gettelman et al. (2019) for an assessment of the stratospheric zonal winds). Here, the jet latitude (speed) is defined as the location (magnitude) of the maximum 850hPa zonal mean zonal wind, determined by a quadratic fit to the winds at the grid point maximum and the two adjacent grid points. During the summer (DJF), LENS exhibited zonal mean westerlies that were around 2ms^{-1} too strong on the poleward side of the jet and around 1ms^{-1} too weak on the equatorward side of the jet (Fig. 2d). This lead to a jet that was slightly too far poleward and slightly too strong (Fig. 2i, red). The strong bias on the poleward side of the jet is still present in CESM2 (Fig. 2c) but is shifted slightly equatorward, such that the CESM2 jet stream is actually stronger than that in LENS, but is in approximately the correct location (Fig. 2i, green and blue).

The representation of the SH winter (JJA) jet stream in LENS was excellent, with the jet latitude and speed being very close to observed and the only bias of note in the zonal mean *ua* being sub-tropical westerlies that were too strong (Fig. 2h). In CESM2, this bias in the sub-tropical westerlies remains, but in addition, a substantial westerly bias of around 2 ms^{-1} is now found in the mid-latitudes (Fig. 2g). This leads to a wintertime jet stream that is stronger than observed, but located at the correct latitude (Fig. 2j green and blue).

Figs. 2i and j demonstrate the prevalance of an equatorward bias in the jet position across both the CMIP5 and CMIP6 models. Curtis et al. (2020) recently argued that, for the annual mean, the equatorward bias in the CMIP6 models was considerably reduced compared to that of CMIP5. For the models and time period considered here, which differ slighly from that in Curtis et al. (2020), we indeed find that the CMIP6 (including BCAM6 and BWACCM6) ensemble mean equatorward bias in the annual mean is only around $1.1°$ compared to around $2.6°$ in CMIP5. However, in the JJA season alone, not considered by Curtis et al. (2020), we still find an ensemble mean equatorward bias in the CMIP6 models of around $3.3°$, which can be compared with a value of around $4.3°$ for the CMIP5 models. Thus, while there are some improvements, a substantial equatorward bias still remains in CMIP6, in JJA in particular, and there are still some models that place the SH westerlies about $16°$ too far equatorward in this season (Fig. 2j). These biases are very large and can be compared with, for example, projected poleward shifts of the jet stream under climate change ranging between 0 and 6° (Barnes & Polvani, 2013; Simpson & Polvani, 2016). LENS had an excellent representation of the SH westerlies in comparison to other CMIP models and in CESM2 the fidelity of the jet position has been maintained but, unfortunately, the SH westerlies are now around 2-3 ms^{-1}

too fast in all seasons (Figs. 2 and S1). A tendency toward zonal mean westerlies that are too strong is common among the CMIP models (Figs. 2i and 2j), although not universal. However, CESM2 lies on the strongest end of the CMIP scale.

A local view of the SH jet stream and storm track is provided in Fig. 3. Here 850hPa ua is used to depict the jet stream (contours) and 850hPa 10-day high pass filtered meridional wind variance ($va'va'$), using a Lanczos filter with 91 weights, is used to depict the storm track (shading). During the summer (Fig. 3a), the SH westerlies and storm track are more zonally symmetric and located closer to the pole than their wintertime counterpart (Fig. 3e) (Hoskins & Hodges, 2005). CESM2 broadly captures the main characteristics of the jet stream and storm track (Fig. 3 b and f) but during the summer, the jet stream exhibits a zonally symmetric westerly bias at the latitude of the Drake Passage (Fig. 3c). This bias is similar to that in LENS (Fig. 3d) but is larger in magnitude and extends farther equatorward. The CMIP6 models exhibit a similar westerly bias around New Zealand and similar easterly biases in the sub-tropical Atlantic and Pacific to those found in CESM2, but they do not uniformly exhibit a similar westerly bias around the Antarctic continent (Fig. A1a).

Substantial local biases in the jet stream also exist in CESM2 in winter (Fig. 3g) and this represents a considerable degradation compared to LENS (Fig. 3h). The westerlies are too strong to the south of Australia (also common to other CMIP6 models (Fig. A1b)) and to the south of South America, leading to the jet stream being less localized in the Indian ocean sector in CESM2 than in observations.

Turning now to the representation of lower tropospheric storm track activity, improvements are seen in all seasons in CESM2 compared to LENS. LENS was characterized by a hemispheric lack of storm track activity, resulting in an unconditional bias in this field (salmon pink component of the LENS bar in Figs. 3 and S2 j and l). An unconditional bias in the sense of a lack of storm track activity is common to the CMIP5 and 6 models (the prevalence of blue circles in Figs. 3 j and l). In CESM2 this unconditional bias has been substantially reduced, leading to an overall reduction in *NMSE* (green and blue bars in Figs. 3 and S2 j and l). In summer, the storm track activity was too weak over the whole of the southern mid-latitudes in LENS (Fig. 3d) and while a weak bias still remains in the ocean basins in CESM2, it is reduced (Fig. 3c). In winter, LENS exhibited a rather dramatic low bias in the lee of the Andes as well as in the Atlantic and Indian Ocean basins - features that are common to other CMIP6 models (Fig. A1d). These biases are substantially reduced in CESM2, although the hemispheric increase in storm track activity has now resulted in the Pacific sector exhibiting too much (Fig. 3g). The role of changes to the orographic drag and blocking schemes in the improvement in the lee of the Andes is discussed further in section 4.3. Too little winter storm track activity off the coast of Antarctica, around the dateline, is a common feature to LENS (Fig. 3h), CESM2 (Fig. 3g) and the other CMIP6 models (Fig. A1d) and Meehl et al. (2019) demonstrated that this bias can be alleviated with increased resolution.

4.2 Northern Hemisphere

The structure of the North Atlantic jet stream and storm track has received considerable attention in prior model intercomparisons, given the importance of this feature for the weather and climate of Western Europe. During winter, the North Atlantic jet stream and storm track are tilted from southwest to northeast and many prior model assessments have demonstrated that models are deficient in the representation of this tilting and instead have an overly zonal jet stream that intercepts the European continent too far south (Doblas-Reyes et al., 1998; Woollings, 2010) along with a lack of storm track activity in the Norwegian sea (Ulbrich et al., 2008; Zappa et al., 2013). It has been argued that this aspect of the circulation is sensitive to resolution (Doblas-Reyes et al.,

1998; Greeves et al., 2007) and the representation of orographic drag and blocking (Pithan et al., 2016; van Niekerk et al., 2017). In addition, metrics of the North Atlantic jet structure averaged over the basin indicate that models often exhibit an Atlantic jet stream that is too far south during the winter (Woollings & Blackburn, 2011; Barnes & Polvani, 2013; Delcambre et al., 2013) and a general tendency toward storm tracks that are too weak in summer and winter, both in terms of cyclone number and intensity (Chang et al., 2012; Zappa et al., 2013).

As demonstrated by Woollings et al. (2010), the climatology of the North Atlantic jet stream in winter is really the average over three preferred states. These states are apparent in the probability distribution function (PDF) of daily jet latitude which indicates three preferred jet locations: the southern most is thought to be associated with the occurrence of Greenland blocking (Hannachi et al., 2012); the central associated with the positive phase of the East Atlantic pattern (Hannachi et al., 2012); and the northern position recently argued to be related to the occurrence of Greenland tip-jet events (White et al., 2019). Both the CMIP3 and CMIP5 models were shown to be deficient in their representation of this trimodal structure, instead exhibiting a jet latitude PDF that is too narrow and too peaked in the central location (Hannachi et al., 2013; Iqbal et al., 2018).

Compared to the Atlantic, previous model intercomparisons have exhibited less consistency in Pacific jet stream biases. In terms of jet location, Barnes and Polvani (2013) showed the CMIP5 models to be rather evenly distributed around the observed location in the annual mean. In the upper troposphere, Delcambre et al. (2013) demonstrated widely varying representation of the jet stream across CMIP3 models, particularly in the jet exit region, but again with a lack of consistency in the nature of the biases.

Given the larger land fraction and prevalence of stationary waves in the NH, we consider metrics for the Pacific and Atlantic jet streams separately. The local jet latitude and speed are defined as the latitude and magnitude of the maximum ua at 850hPa, between 20°N and 65°N, at each longitude, determined by a quadratic fit to the grid point maximum and the two adjacent grid points. For the Atlantic basin, the jet latitude and speed are defined as the mean of the local jet latitudes and speeds over the grid points between 60°W and 10°W but we note that the same conclusions can be drawn for the latitude and speed of the zonal wind averaged over the Atlantic sector first. The Atlantic jet tilt, in degrees latitude, is defined as the angle, relative to the zonal direction, of the best fitting straight line to the local jet latitudes between 60°W and 10°W. For the Pacific, these calculations are performed over 150°E to 130°W.

For the North Atlantic during winter, we will assess the PDF of daily jet latitude as defined by Woollings (2010). The location of the jet maximum between 15°N and 75°N is obtained for the 10 day low pass filtered (lanczos filtered with 91 weights) daily 850hPa ua, averaged between 60°W and the Greenwich Meridian. This jet latitude timeseries is then deseasonalized by removing the annual mean and the first three harmonics of the seasonal cycle. The jet latitude PDF is presented using the kernel estimate of Silverman (1981) (their equation 1) with the smoothing parameter $h=1.06\sigma n^{-1/5}$, where σ is the standard deviation and n is the sample size, as in Woollings (2010).

During winter, the ERA5 North Atlantic jet stream is located at 46.6°N (Fig. 4b) with a mean speed of 9.8ms^{-1} (Fig. 4c) and is tilted by ∼16° latitude (Fig. 4d), although both LENS and the twentieth century reanalyses indicate substantial uncertainty on these metrics with a 36 year record. Nevertheless, Figs. 4a-d demonstrate that even given the sampling uncertainty, many of the CMIP5 and CMIP6 models exhibit a North Atlantic jet stream that is located too far south, is too fast and is lacking in the southwest to northeast tilt, in agreement with the aforementioned previous studies. The ensemble spread in BCAM6 encompasses the observed value in both jet latitude and tilt and it intercepts the continent at close to the correct location (over the UK), while the majority of other

CMIP models intersect the continent over France (Fig. 4a). The CESM2 wintertime North Atlantic jet is, however, too fast by about 2 ms^{-1} in all configurations (Fig. 4c). Interestingly, the representation of the North Atlantic jet stream is degraded in WACCM6 compared to CAM6, in that it does not exhibit the correct southwest to northeast tilt in the east Atlantic, meaning that the jet stream intersects the continent at lower latitudes over France, like many of the other CMIP5 and 6 models (Fig. 4a). Given that this difference between the high top and low top configuration is consistent between the coupled and uncoupled configurations, it is likely a robust feature of the difference between CAM6 and WACCM6, but remains to be understood.

Compared to winter, the summer North Atlantic jet stream is situated further north, is weaker and is less tilted (Figs. 4e-h). The CMIP5 and CMIP6 models are relatively evenly distributed around the observed values with no clear tendency toward a bias of one sign or the other (Fig. 4f-h, gray points). LENS simulated these characteristics of the North Atlantic jet extremely well with its range encompassing ERA5 in all three metrics (Figs. 4f-h, red range). While CESM2 captures the summertime jet tilt correctly, it now places the jet stream too far North and the jet speed is about 1ms^{-1} too fast. In contrast to the wintertime, the summertime jet structures in CAM6 and WACCM6 are very similar to each other.

The PDF of DJF daily jet latitude for the Atlantic jet is shown in Fig. 4i. The reanalyses exhibit three preferred jet latitude locations: the southern location about 12° south of the mean jet; the central location, about 3° south of the mean jet; and the northern location about 8° north of the mean jet. Interestingly, both ERA5 and JRA55 exhibit an additional local maximum in the PDF around 2° north of the mean jet - a feature not present in ERA-Interim or MERRA2. To assess the representation of this PDF for each of the models, the probability of the jet being located in the 5 degree latitude bands centered on the southern, central and northern ERA5 PDF peaks (gray shading in Fig. 4i) is shown in Figs. 4 j-l. The probability of occurrence in the central locations is well represented by CESM2, LENS and many of the CMIP5 and CMIP6 models. The probability of occurrence in the northern location is also represented well in CESM2 and LENS while many of the CMIP5 and CMIP6 models exhibit a reduced probability in this region, consistent with Hannachi et al. (2012)'s conclusions that the PDF is too narrow. The two features of concern for CESM2's representation of the jet latitude PDF are that there is a reduced probability compared to observations of the jet being situated in the southern location, with a compensating enhanced probability of it sitting between the central and northern locations (Fig. 4i). The reduced probability of occurrence in the southern location is a feature that is common to many CMIP models (Fig. 4j) and LENS (Kwon et al., 2018) and is consistent with the reduced number of Greenland blocking events compared to observations. This aspect looks to be slightly improved in CESM2 compared to LENS, which is consistent with an improvement in Greenland blocking to be discussed in section 6.4.1. An interesting difference between CAM6 and WACCM6 is that, relative to CAM6, WACCM6 exhibits an increased probability of the jet being situated in the central latitudes and a reduced probability of it being situated in the northern latitudes (compare green and blue in Fig. 4i). This is a robust difference present in both the coupled and uncoupled runs, but it's not clear whether WACCM6 or CAM6 is closer to observed.

For the Pacific jet stream (Fig. 5), during winter, all three jet metrics are rather well simulated in CESM2 (Figs. 5 b-d). Many of the CMIP5 and 6 models and LENS have the Pacific jet stream placed too far equatorward, but this is not true in coupled CESM2 and the uncoupled CESM2 runs actually exhibit a Pacific jet stream that is too far poleward (Fig. 5b). During the summer, the Pacific jet is placed too far poleward in all CESM2 configurations - a problem that was also present in LENS (Fig. 5f). In addition, the Pacific jet stream has now become a bit too strong in CESM2, while it's speed was well represented in LENS (Fig. 5g).

Taking a broader view of the spatial structure of the jet streams and storm tracks by considering 850hPa zonal wind and 10-day high pass filtered eddy meridional wind variance in Fig. 6, it is clear that the climatology during the wintertime is greatly improved in both aspects in CESM2. The large biases in 850hPa ua that were present in the Pacific in LENS leading to a jet stream that was placed too far equatorward are now gone and the biases that extended across the Atlantic basin are reduced (Fig. 6c compared to d). The only bias of note in 850hPa ua that remains, is the westerlies that are too strong over western Europe and an accompanying easterly bias over North Africa (Fig. 6c), which is a feature common to the majority of CMIP6 models (Fig. A1e). These climatological biases in CESM2 of around 3m/s over Europe and North Africa are larger than the average projected end-of-century changes in this region by the end of the century which are of the order 1m/s (Woollings & Blackburn, 2011, their Figure 1a). Nevertheless, consideration of the overall *NMSE* for DJF 850hPa ua (Fig. 6i) makes clear that CESM2 is a top ranking model in this aspect and is improved compared to LENS.

Massive improvements are also found in the representation of the NH winter storm tracks (Figs. 6c and d). As measured by the 850hPa 10-day high pass filtered eddy meridional wind variance ($va'va'$), LENS exhibited storm tracks in the Atlantic, Pacific and the lee of the Rockies that were too weak - in some regions by a very large fraction (>50%). A tendency toward weaker storm tracks than observed in the lee of the Rockies, over the North Atlantic and in the central Pacific is also common among the CMIP6 models (Fig. A1g) and when a large unconditional bias is present in the CMIP models, it is typically negative i.e., the storm tracks are too weak (the prevalence of blue circles in Fig. 6j). The biases that were present in LENS are now greatly reduced in CESM2 and this appears as a substantial reduction in *NMSE* arising from both an elimination of the unconditional bias and a reduction in the phase error (Fig. 6j), leaving CESM2 as one of the highest ranking CMIP6 models. The role of changes in the orographic drag formulation in the improvements in the lee of the Rockies is discussed in section 4.3 below.

During the summertime, there has been a degradation in the representation of the overall 850hPa ua field in CESM2 compared to LENS, largely due to strengthened westerlies around the Arctic circle. This degrades even further in the uncoupled simulations. Despite the degradation in ua, large improvements are seen in the NH storm tracks in JJA, again through the elimination of the negative unconditional bias that was present in LENS and is found in many other CMIP5 and 6 models (light gray(salmon pink) component of the CMIP(LENS) bars in Fig. 6l).

4.3 The influence of new orographic schemes on storm track activity in the lee of large mountain ranges

As discussed above, in the lee of both the Andes in the SH and the Rockies in the NH, there are large improvements in the representation of lower tropospheric meridional wind variance on synoptic timescales. These improvements are more notable in the winter season of each hemisphere and are further highlighted in Fig. 7 for the Rockies during DJF (panels a-c) and for the Andes during JJA (panels i-k). The large increase in meridional wind variance in the lee of the mountains in the coupled simulations is also present in the uncoupled difference between FCAM6* and FCAM5 i.e., where the CAM6 physics package is replaced with CAM5 (Figs. 7d versus e and l versus m). A large fraction of the increase in storm track activity in the lee of these mountain ranges can be explained by the sum of the influences of the new Turbulent Orographic Form Drag (TOFD) and Mesoscale Orographic Blocking (MOB) schemes. See panels (f) and (n) for their sum and (g-h) and (o-p) for their individual contributions. The TOFD scheme is by far the dominant influence in the lee of the Rockies. In the lee of the Andes, both play a role, but the sum of their individual influences falls short of the total change found in going from CAM5 to CAM6 physics. We do not investigate the additional influences on the downstream storm tracks here as the change in storm track activity over the ocean basins

to the east is more muted in the CAM6*−CAM5 difference than in the BCAM6−LENS difference, but it is something that would be interesting to address in coupled simulations in a subsequent study.

5 Stationary waves and the global divergent circulation

5.1 Stationary waves

Stationary waves, or zonal asymmetries, in the atmospheric circulation arise through interactions between the zonal mean flow, orography, diabatic heating and transient vorticity and divergence fluxes (e.g. H. Wang & Ting, 1999; Held et al., 2002). They play an important role in determining regional climate (e.g. Broccoli & Manabe, 1992; Rodwell & Hoskins, 2000) and in projected future regional climate change (e.g. Seager et al., 2014; Shaw & Voigt, 2015; Simpson et al., 2016; Wills et al., 2019) and are, therefore, an important aspect to simulate correctly. Notable biases that have been discussed in previous model intercomparison studies are the inability of models to accurately capture the trough/ridge pattern over the North Atlantic in winter (Boyle, 2006) and the inability of models to capture the localized anticyclonic circulation to the south of New Zealand in summer (Simpson et al., 2013). These issues are related to the overly zonal jet stream and storm track in the North Atlantic and the zonally extended SH jet stream discussed above. Here, we assess the representation of the climatological stationary waves using 300hPa eddy stream function (ψ^*), but the extra-tropical features have an equivalent barotropic structure and similar conclusions can be drawn throughout the depth of the troposphere.

A comparison of Figs. 8b and c indicates substantial improvements in the representation of DJF NH stationary waves in CESM2 compared to LENS. The large biases in LENS over the sub-tropical and mid-latitude Atlantic and Pacific oceans have been substantially reduced in accordance with the improvements in the jet streams discussed above. The primary bias that remains is the anticyclonic circulation centered over the Mediterranean and cyclonic anomalies to the south and north of this. These features are common among the CMIP6 models (Fig. A1i) and are associated with the westerly (easterly) biases over Europe (North Africa) discussed above. We also note that the high latitude stationary wavenumber 1 bias present in CCSM4 discussed in Shaw et al. (2014) has been alleviated in both CESM1 and CESM2 (not shown). In terms of the NH winter *NMSE*, CESM2 is one of the highest ranking models in the CMIP6 archive and is improved over LENS through reductions in both the conditional bias and the phase error (Fig. 8g).

During JJA, the representation of the stationary waves in the Pacific-North American sector is quite comparable to that in LENS (Fig. 8e compared to f). The positive ψ^* bias that was present over the Middle East and North Africa in LENS has been reduced, leading to an overall reduction in the *NMSE* (Fig. 8h) and making CESM2 one of the top ranking models in this regard in the CMIP6 archive. However, compared to the coupled simulations, the representation of the NH summer stationary waves is degraded in the simulations with prescribed observed SSTs (hatched bars in Fig. 8h) and the role of coupling versus different underlying SSTs in this will be discussed in section 5.3 below.

Unfortunately, the representation of SH stationary waves has not fared so well under the CESM2 developments. In the summer, the SH stationary wave representation in CESM2 is rather comparable to that in LENS (Figs. 8 b, c and i). The SH winter stationary waves were represented extremely well in LENS and it ranked among the top of all CMIP5 and CMIP6 models (Fig. 8j). In CESM2, along with the degradation of the SH westerlies discussed above, the SH winter stationary waves have degraded across the whole of the mid-latitudes with, most notably, a large cyclonic bias to the south of Aus-

tralia. This has resulted in a substantial increase in *NMSE* for SH stationary waves in JJA, although CESM2 still sits roughly in the middle of the CMIP6 range in this metric (Fig. 8j).

5.2 The Global Divergent Circulation

The global divergent circulation is intrinsically connected to the hydroclimate of the tropics and the forcing of extra-tropical stationary waves (Sardeshmukh & Hoskins, 1988) and is, therefore, a feature that should be accurately simulated in order to give a reasonable representation of the climate system, but it is more challenging to observationally constrain than the extra-tropical rotational flow.

We begin our analysis of the divergent circulation by assessment of bulk measures of the low latitude, zonally symmetric, mean meridional circulation, or Hadley circulation, given by

$$\Psi = \frac{2\pi cos\phi}{g} \int_0^p vadp, \qquad (5)$$

where va here refers to the zonal mean meridional wind, p refers to pressure level, ϕ to latitude and g to the accelereration due to gravity. Figs, 9a-b and e-f indicate that CESM2 depicts the summer and winter Hadley cells with reasonable fidelity. A noticeable difference is present in DJF at low levels in the equatorial region where the clockwise overturning circulation extents too far South - a feature that was also present in LENS (Fig. 9c and d). In JJA, the anti-clockwise cross equatorial Hadley cell is now too strong (Fig. 9g), whereas it was slightly too weak in LENS (Fig. 9h). In the equinox seasons, there are notable biases in the mean meridional circulation on the equator which reflect the circulation of the spring-time hemisphere encroaching too much into the lower latitudes (Figs. S7 and S8).

Figs. 9 i and j (top) show the Hadley cell extent defined as the zero-crossing latitude of the Hadley cell streamfunction in the winter hemisphere. This is obtained at 500hPa by the linear interpolation between the two grid points either side of the zero crossing (Adam et al., 2018). The scatter of the reanalysis points illustrates the challenge in observationally constraining this measure, but overall CESM2 lies within the range of observational uncertainty. In both DJF and JJA, it is extremely close to the ERA5 value. Hadley cell extent varies by about a $\sim5°$ latitude range across the CMIP5 and CMIP6 models in each season, but they are evenly distributed around the observed value, with no systematic tendency toward a bias of one sign or the other.

We define the Hadley cell intensity as the maximum magnitude of the the mean meridional stream function at 500hPa, determined by a quadratic fit to the grid point maximum of the absolute value of stream function and the two adjacent grid points. The CMIP models exhibit wide ranging values for this metric of Hadley cell intensity with a slight propensity for an overly strong Hadley circulation (Figs. 9i and j, bottom). Similar degrees of inter-model spread have been found in older model intercomparisons and Caballero (2008) related this to the varied representation of extra-tropical eddy driving. The Hadley cell intensity is well represented in CESM2, and lies within the reanalysis ranges in both seasons.

We provide a more local view of the upper level divergent flow in Fig. 10 with the 200hPa velocity potential (χ). Velocity potential is proportional to the inverse Laplacian of divergence and is related to ua and va as follows:

$$ua = -\frac{\partial \chi}{\partial x} \quad va = -\frac{\partial \chi}{\partial y}, \qquad (6)$$

such that the divergent winds flow perpendicular to isolines of χ, out of regions of negative χ and into regions of positive χ. During DJF, the global structure of χ depicts divergence in the western tropical Pacific and south Pacific convergence zone and conver-

gence over the sub-tropical Atlantic, Mediterranean and eastern sides of the southern sub-tropical ocean basins (Fig. 10a). In JJA, the maximum divergence is present over Asia and the north tropical Pacific, with convergence in the southern sub-tropical Atlantic (Fig. 10d).

During DJF, LENS exhibited too much divergence over the western Indian ocean (Fig. 10c), which is a common feature among CMIP6 models (Fig. A1k) and was associated with too much precipitation in that region. In CESM2, the excess divergence is reduced and shifted further west and accompanies a reduction in the excess precipitation bias (Danabasoglu et al., 2019). Overall, the errors in the 200hPa velocity potential field have been reduced in DJF, making CESM2 one of the top ranking models in this metric (Fig. 10g).

During JJA, the representation of 200hPa velocity potential is also substantially improved in CESM2 compared to LENS (Figs. 10e and f) with the large biases that were present over the Indian Ocean, East Asia, tropical Atlantic and Amazonia now greatly reduced, in association with precipitation improvements in these regions (Meehl et al., 2020). CESM2 is a top ranking model in the JJA global divergent flow field and rather remarkably, the *NMSE* in CESM2 is not that much larger than for ERA-Interim relative to ERA5 (Fig. 10h). That being said, the difference between ERA5 and ERA-Interim is not in the same location as the biases found in CESM2, so there are still improvements that can be made. Furthermore, as with the JJA NH stationary waves, Fig. 10h makes clear that the simulation of the global divergent circulation is degraded in uncoupled simulations with prescribed observed SSTs - a feature that will now be discussed in more detail.

5.3 The role of ocean-atmosphere coupling and SST differences in the difference between coupled and uncoupled simulations.

Compared to the coupled simulations, the representation of the NH stationary waves and global divergent circulation in JJA is degraded when observed SSTs are prescribed in both CAM6 and WACCM6 (Figs. 8h and 10h). There are two possible reasons for this. One is that the lack of ocean-atmosphere coupling in the prescribed SST simulations could be responsible for the degradation. If that were the case, then we should not be concerned about this degradation as the presence of coupling is more realistic. The other possibility is that the different SSTs in the real world, compared to the model, are responsible. If this is the case, then it would indicate that compensating biases, one of which being biases in the SST field, are responsible for the improvements in the coupled simulation over the uncoupled simulations.

To assess which of these is true, a comparison with the simulation FCAM6MOD, where SSTs from the model are prescribed (Table 2), is provided in Fig. 11. This reveals that the biases in both χ and NH ψ^* in FCAM6MOD are similar to those in BCAM6 (compare Fig. 11d with 8e and Fig. 11g with 10e) and smaller than those in FCAM6 (Figs. 11c and f). FCAM6 exhibits larger biases than both FCAM6MOD and BCAM6 in χ over the Atlantic and Indian Ocean sectors (Fig. 11f) and in ψ^* over much of North Africa and Eurasia. The implication of this is that the degradation in the uncoupled compared to coupled simulations arises through differences in the SSTs and not the lack of coupling, indicating that compensating errors are contributing to the fidelity of the coupled simulation. Compared to the modelled SSTs, the observed SSTs are cooler in the tropical Atlantic (Fig. 11a) which is likely responsible for the reduced divergence in that region in FCAM6 and compensating reduced convergence over the Middle East and Asia (Fig. 11h) and the degradation in the stationary wave representation over North Africa and Eurasia (Fig. 11e).

6 Extra-tropical variability: SAM, NAM, NAO and Blocking

In addition to representing the climatological features of the large scale atmospheric circulation with fidelity, ESM's must also accurately represent the higher frequency variability of the climate system. In this realm, there are many features that could be assessed. Here we focus on the dominant modes of extra-tropical atmospheric variability along with the representation of atmospheric blocking, given the attention that has been paid to these features in previous model intercomparisons. Note that we have also presented a climatological view of extra-tropical storm track statistics and North Atlantic jet variability in section 4 above, which could also fall under this category of extra-tropical variability.

6.1 The Southern Annular Mode (SAM)

The SAM, also known as the Antarctic Oscillation, is the dominant mode of variability in the SH extra-tropical circulation (D. Gong & Wang, 1999; Thompson & Wallace, 2000). It can be defined using zonal wind, geopotential height or sea level pressure on daily to seasonal timescales and is most commonly identified as the first Empirical Orthogonal Function (EOF) of variability in these fields. In the troposphere, the positive phase of the SAM represents a poleward shifting of the eddy-driven mid-latitude westerlies, while in the stratosphere it represents a strengthening of the stratospheric polar vortex.

A model's representation of the tropospheric SAM is of particular interest, not only because of its influence on regional surface temperature and hydroclimate variability (Gillett et al., 2006), but also because the SH mid-latitude circulation response to many forcings, such as ENSO (L'Heureux & Thompson, 2006), ozone depletion (Thompson & Solomon, 2002) and increasing GHG concentrations (Miller et al., 2006; Gillett & Fyfe, 2013), projects strongly onto the SAM. The dominance of the SAM in both internal variability and the response to forcings in the troposphere is thought to arise because of a positive eddy-mean flow feedback that occurs in response to SAM-like circulation anomalies (Lorenz & Hartmann, 2001), although this interpretation has recently been questioned (Byrne et al., 2016).

One metric of SAM variability that was thought to provide an indication of the strength of eddy-mean flow feedbacks is its persistence timescale (Gerber & Vallis, 2007), although this is complicated by the fact that additional forcings e.g., stratospheric variability, also impart additional persistence on the SAM (Baldwin et al., 2003; Simpson et al., 2011). Prior model intercomparison studies have demonstrated the tendency for models to exhibit an overly persistent SAM (Gerber, Polvani, & Ancukiewicz, 2008; Gerber et al., 2010; Kidston & Gerber, 2010; Son et al., 2010), raising concern that the eddy-mean flow feedbacks in such models may be too strong and, as a result, they may produce a response to external forcings that is too large. Indeed, in support of this viewpoint, Kidston and Gerber (2010) argued that CMIP3 models with a more persistent SAM also exhibited a larger response to GHG forcing. However, Simpson and Polvani (2016) have since shown with a larger number of CMIP5 models and through consideration of individual seasons, that there is no significant relationship between a models SAM timescale and its response to external forcings. So, while there is good theoretical reasoning to expect the SAM timescale to have bearing on a models response to external forcing (Leith, 1975), and there is ample evidence for such a correspondence from idealized modelling studies (e.g. Gerber, Voronin, & Polvani, 2008), evidence for the connection between SAM persistence and the circulation response to forcings in the more comprehensive system is currently lacking.

Here, we follow Gerber et al. (2010) and define the SAM for the full year as the first EOF of daily zonal mean geopotential height after subtracting the global mean, deseasonalizing and linearly detrending. The EOF is calculated on each vertical level separately, using cosine weighted data from the equator to 90°S and the SAM index is the

accompanying Principal Component (PC) time series. To determine the SAM timescale, the lagged autocorrelation function of the SAM index is first determined for each day of the year according to Eq 1 of Simpson et al. (2011). This is then smoothed over a 181 day window using a Gaussian filter with a full width at half maximum of around 42 days and the SAM timescale is the e-folding timescale of a least squares exponential fit to this smoothed autocorrelation function out to a lag of 50-days.

Fig. 12a presents the zonal mean structure of the SAM on selected pressure levels for BCAM6 and BWACCM6, weighted by cosine of latitude so that it represents a mass displacement. These height anomalies are in geostrophic balance with a poleward shifting of the mid-latitude westerlies in the troposphere and a strengthening of the polar vortex in the stratosphere. Following Gerber et al. (2010), in Fig. 12b we characterize the SAM structure by the latitude of the node of the dipolar zg anomalies i.e., the mid-latitude zero crossing point as illustrated in Fig. 12a. In Fig. 12c we show the latitude weighted, root mean square amplitude of the annular mode structure. These are shown as a function of height for each CESM2 simulation and the reanalysis and only at the 500hPa for all other simulations.

CESM2 captures the amplitude of the SAM very well, with the reanalysis amplitude sitting within the range of the 10 BCAM6 members for all levels except 10hPa, where the amplitude of reanalysis variability is slightly lower than that found in the model (Fig. 12c). Most CMIP models also simulate a SAM amplitude at 500hPa that is close to observed. The location of the SAM node is well captured in the troposphere in CESM2 (Fig. 12b), which is in contrast to many CMIP models where the node is placed too far equatorward in association with the equatorward bias in their climatological jet location (section 4.1). In the lower stratosphere, the location of the SAM node in CESM2 is displaced slightly equatorward of that found in reanalysis.

The latitude-longitude structure of geopotential height variability that accompanies the zonal mean SAM can be assessed in Figs. 12d-g which show the regression of 500hPa geopotential height (after deseasonalizing and detrending) onto the SAM index. While the SAM is predominantly zonally symmetric, some zonal asymmetries do exist in its structure and these are more pronounced in the reanalysis than in CESM2 (Fig. 12d versus e). In the reanalysis, a localized intensification of the Antarctic low occurs over the Amundsen-Bellingshausen sea region which leads to poleward flow toward the west of the Antarctic peninsula and contributes to warmer surface temperatures over West Antarctica during the postive phase of the SAM as well as regional sea ice anomalies (Lefebvre et al., 2004; Sen Gupta & England, 2006). This feature is largely absent in CESM2 as evidenced by the large positive height bias off West Antarctica in Fig. 12f, but it was well reproduced in LENS (Fig. 12g). In addition, CESM2 is lacking in the local intensifications of the positive height anomalies to the south east of New Zealand, off the east coast of South America and in the Southern Indian ocean. Similar biases are common among the CMIP6 models (Fig. A2a) but this was rather well represented in LENS (Fig. 12g). Nevertheless, when considering the *NMSE* of the 500hPa SAM structure (Fig. 12k), despite being degraded slightly compared to LENS, CESM2 still ranks highly among the CMIP5 and CMIP6 models.

CESM2 accurately captures the enhanced SAM persistence during spring in the stratosphere and the associated enhanced SAM persistence in the spring and early summer in the troposphere (Figs. 12h-j), which arise due to variability in the timing of the polar vortex breakdown and its downward influence (Baldwin et al., 2003; Simpson et al., 2011; Byrne & Shepherd, 2018). However, as is commonly found in models (Gerber et al., 2010), the enhanced tropospheric persistence extends too late into January and February. Given the importance of stratospheric variability in SAM persistence, rather than comparing with LENS, we compare BCAM6 with BWACCM6 in Fig. 12j. The SAM persistence in BCAM6 is rather comparable to that in BWACCM6 and the greater springtime stratospheric persistence in BWACCM6 is not significant. A comparison across mod-

els of the DJF averaged SAM persistence at 500hPa is provided in Fig. 12l. This is the season when prior studies have demonstrated an overly persistent SAM in models, while during JJA, all models compare well with observations (not shown). Many of the CMIP5 and CMIP6 models exhibit SAM persistence that is greater than observed, although LENS indicates that the sampling uncertainty is large. At least from the data currently available, none of the CMIP6 models exhibit extreme SAM persistence like that found in a few CMIP5 models. In particular, the IPSL model which was extremely persistent in CMIP5 is now improved in CMIP6. All CESM2 configurations have a reasonable representation of SAM persistence.

Overall, the structure and persistence of the SAM is well represented in CESM2, but there are some notable degradations in the representation of zonal asymmetries in the SAM structure around the Antarctic continent which may have implications for the representation of surface temperature and sea ice variability.

6.2 The Northern Annular Mode (NAM)

The NAM, also known as the Arctic Oscillation, is the NH equivalent to the SAM and is the dominant mode of NH zonal mean extra-tropical circulation variability. It is accompanied by surface signatures in temperature and precipitation and forced circulation responses are often found to project onto the NAM (Thompson & Wallace, 1998; Gillett & Fyfe, 2013; Miller et al., 2006). Previous model intercomparison studies that assessed the monthly NAM have shown a systematic tendency for models to underestimate the amplitude of the centers of action in the Atlantic, while overestimating the centers of action in the Pacific, and to overestimate the negative height/surface pressure anomaly over Siberia (Stoner et al., 2009; Zuo et al., 2013; H. Gong et al., 2017). In addition, Miller et al. (2006) argue that the NAM represents too large a fraction of temporal variability in the NH in models.

We present an assessment of daily NAM variability using exactly the same methodologies as described above for the SAM, but calculating the EOF using data from the equator to 90°N. The overall zonal mean structure of the NAM is well represented in CESM2 throughout the troposphere and lower stratosphere (Fig. 13a), with the NAM node being very well represented in WACCM6 but slightly too far poleward(equatorward) in the troposphere(stratosphere) in CAM6 (Fig. 13b). The tropospheric NAM amplitude is slightly too large in all configurations and an interesting difference between CAM6 and WACCM6 appears higher up in the stratosphere, where the amplitude of NAM variability in WACCM6 is reduced, and further from the observations, compared to that in CAM6 (Fig. 13c). This is unexpected given the enhanced representation of the stratosphere in WACCM6 compared to CAM6. Since the model top in CAM6 is rather low (∼2.26hPa) and the layers below that are influenced by the sponge-layer, chances are high that this improved representation of the NAM amplitude in CAM6 compared to WACCM6 is due to spurious reasons, although it remains to be fully understood.

The horizontal 500hPa NAM structure in CESM2 broadly resembles that in the observations (Fig. 13e versus d). As in CESM1 (Fig. 13g) and other CMIP6 (Fig. A2b) and CMIP5 models (Zuo et al., 2013), there is a significant overestimation of the amplitude of the negative height anomalies over Northern Russia (Fig. 13f). Compared to LENS, the representation of the NAM structure in the western Atlantic is substantially improved in CESM2, but it is degraded in the Pacific where the amplitude of the positive height anomalies there in CESM2 are now larger than observed (Fig. 13g versus f). This overestimation of the amplitude of the Pacific center of action is a similar bias to that found in many other CMIP6 models (Fig. A2b). Overall, the structure of the NAM, as quantified by the *NMSE*, appears to be slightly improved in CESM2 compared to LENS since a number of the CESM2 members have a lower *NMSE* than the LENS range (Fig. 13k).

Both the low top and high top CESM2 configurations exhibit enhanced stratospheric NAM persistence and the concomitant increase in tropospheric NAM persistence during the winter, when longer timescale stratospheric variability plays an important role and imparts persistence to the troposphere (Baldwin et al., 2003; Simpson et al., 2011). The ensemble mean NAM persistence suggests enhanced timescales in BCAM6 compared to BWACCM6 during JFM. However, a closer inspection of the individual ensemble members during DJF reveals a large sampling uncertainty in this metric (Fig. 13l) and indicates that this difference is likely not significant. The spread across the BCAM6 members is also as large as the CMIP5 and CMIP6 inter-model spreads, again demonstrating the substantial uncertainty on this metric due to internal variability and questioning its usefulness for model validation.

6.3 The North Atlantic Oscillation (NAO)

The NAO is the dominant mode of variability in the atmospheric circulation of the North Atlantic sector. It is characterized by fluctuations in the sea level pressure difference between the Iceland low and the Azores high and concomitant variations in the strength and location of the North Atlantic jet stream, with implications for surface weather in the North Atlantic sector (Hurrell, 1995; Bladè et al., 2012). The NAO is highly correlated with the Arctic Oscillation, or NAM, but reflects variability that is more localized to the North Atlantic sector and is argued by Ambaum et al. (2001) to be a more physically relevant mode of variability for the NH than the zonally symmetric NAM. To minimize the duplication of information from the NAM calculation above, here we assess the representation of the NAO on lower frequencies using monthly data, although similar conclusions can be drawn for the NAO calculated using daily data (not shown).

In general, prior assessments of the representation of the NAO have indicated that models are capable of simulating the structure of the wintertime NAO, reasonably well (Stephenson & Pavan, 2003; Cohen et al., 2005; McHugh & Rogers, 2005; Stephenson et al., 2006; Stoner et al., 2009). The same is true of the summer NAO, although Bladè et al. (2012) emphasize inadequacies in the representation of associated precipitation anomalies over the Mediterranean, with the amplitude of the precipitation signal being too weak, especially in the east.

While models can reasonably well capture the structure of the NAO and interannual timescale variability, a number of recent studies have indicated that models underestimate the amplitude of multi-decadal NAO and jet stream variability (Kravtsov, 2017; X. Wang et al., 2017; Kim et al., 2018; Simpson et al., 2018). Given that our focus is on the representation of atmospheric circulation as compared to satellite-era reanalyses, we do not address the issue of multi-decadal variability here.

We define the NAO as the first EOF of monthly mean sea level pressure variability, area weighted, over the North Atlantic domain (90°W-40°W, 20°N-80°N) (Hurrell, 1995). Monthly *slp* data were deseasonalized and detrended and the seasonal EOF calculated for the concatenated time series of the individual component months of the season for all years. The PC time series and EOF patterns are constructed such that the PC time series has standard deviation of 1 and the EOF has *slp* units. For CMIP, the NAO pattern and variances are calculated separately for each ensemble member and then averaged. For the models, to account for the possibility of the NAO not appearing as the dominant mode of variability, the first three EOFs are calculated and the one that has the highest spatial pattern correlation with the observed NAO is considered to be the NAO. Only a few model members have the NAO as the second EOF.

The structure of the wintertime NAO is relatively well represented in CESM2 (Figs. 14b versus a). CESM1 exhibited an Azores high anomaly that was too weak (Fig. 14d) and this is now improved in CESM2, although the magnitude of the accompanying Pacific SLP anomaly is now larger (Fig. 14c). A salient feature of the *NMSE* intercompar-

ison (Fig. 14e) is the large uncertainty in this metric across ensemble members of the LENS (red range) and across 35 year overlapping segments of the twentieth century reanalyses (black ranges). These ranges encompass almost the whole of the intermodel spread, which renders any intercomparison somewhat meaningless as the sampling uncertainty on the real world and across members of an individual model are of comparable magnitude to the inter-model spread. One bias feature that is common to all CMIP6 models (Fig. A2c) and CESM2 is a positive SLP anomaly centered over the British Isles and Scandinavia - the region where the North Atlantic jet stream is too zonally extended. The fraction of SLP variance explained by the winter NAO is well represented in CESM2 and most models, given the sampling uncertainty range (Fig. 14f).

The summertime NAO is also well represented in CESM2 (Figs. 14g and h) with some minor biases that are rather similar to those in LENS. The sampling uncertainty on the *NMSE* is smaller, but still covers a substantial fraction of the inter-model spread and CESM2 performs reasonably well and within approximately the same range as displayed by LENS (Fig. 14k). The percent variance explained by the summer NAO is well represented in CESM2 and the majority of CMIP5 and 6 models (Fig. 14l).

6.4 Blocking

Blocking refers to quasi-stationary circulation anomalies that block the mid-latitude westerly flow for several days or more (see e.g., Woollings et al. (2018) for a recent review). Blocks are associated with extreme cold events in winter and heat waves in summer, through their effects on thermal advection and radiative fluxes (Trigo et al., 2004; Sillmann et al., 2011; Buehler et al., 2011; Pfahl & Wernli, 2012; Brunner et al., 2018). Accurate representation of atmospheric blocking is, therefore, necessary for the accurate simulation of extreme events and their projected future changes, yet blocking statistics are often misrepresented in models.

Many blocking indices exist (see Barnes et al. (2013) and Woollings et al. (2018) for example intercomparisons of a variety of blocking indices). Here, we adopt the two-dimensional procedure of Masato et al. (2013a). This method identifies persistent reversals of the meridional gradient of the 500hPa geopotential height field and was used in CMIP5 comparisons in both the NH (Masato et al., 2013b) and SH (Patterson et al., 2019).

Daily 500hPa zg data on a 2° longitude×latitude grid are used to calculate a blocking index B, as follows. At each grid point between 25° and 75° North or South, the following integrals are evaluated

$$Z^P = \frac{2}{\Delta\phi} \int_{\phi_o}^{\phi+\Delta\phi/2.} (zg)d\phi; \quad Z^E = \frac{2}{\Delta\phi} \int_{\phi_o-\Delta\phi/2.}^{\phi_o} (zg)d\phi \qquad (7)$$

with $\Delta\phi$=30° latitude i.e., Z^P and Z^E represent the integrals of geopotential height over the 15° latitude range poleward and equatorward of the grid point latitude ϕ_o, respectively. The blocking index is then given by $B = Z^P - Z^E$, such that when a large scale reversal of the meridional gradient of zg occurs, $B > 0$. For each day, local positive maxima in B are identified within the latitude range 40° to 70°. Each local gridpoint maxima and all adjacent positive values of B that are contiguous to that grid point are considered to be part of the same blocking event. For each day, other than the first day of the season, if a local maximum in B lies within ±18° longitude and ±14° latitude of a blocking center identified on the previous day, then it is considered to be a continuation of that previous day's blocking event. An event is considered to end if there is no local positive B maximum within ±27° longitude and ±20° latitude of the local maximum of the first day of that block. In this way, at each grid point the number of days where $B > 0$ in association with persistent blocking events of a given duration can be quantified. The metric we use for evaluation is the overall fraction of days in the time series that are blocked, expressed as a percentage. To identify the blocking climatologies of the

NH, we consider only blocking events that last 5 days or more. For the SH, we consider blocking events that last 4 days or more, to account for the greater transience of the SH (Berrisford et al., 2007). An assessment of NH blocking using slightly different metrics can also found in Gettelman et al. (2019), their Figure 15 and Benedict et al. (2019).

Since the blocking index used here relies on the absolute threshold that the meridional gradient of zg must reverse, both climatological biases in the mean zg gradient and biases in the representation of the synoptic variability that gives rise to blocks can contribute to a bias in blocking frequency. The implications for the representation of surface climate variability could be rather different depending on which of these dominate e.g., someone looking to use the model to investigate heat waves associated with blocking events will be less concerned if a low bias in blocking occurs because the climatological geopotential gradient is too steep, so it is difficult to reverse, as opposed to if it were due to the model being unable to produce persistent anticyclones. To provide an indication of the relative importance of each of these contributions to the overall bias, we adopt the procedure of Scaife et al. (2010). The geopotential height field is divided up into the seasonal mean climatology and the deviations therefrom for both the model and ERA5 i.e.,

$$zg_{mod}(t) = \overline{zg}_{mod} + zg'_{mod}(t) \tag{8}$$

$$zg_{era}(t) = \overline{zg}_{era} + zg'_{era}(t), \tag{9}$$

where $\overline{(.)}$ refers to the seasonally averaged climatology from 1979-2014, $(.)'$ refers to deviations therefrom, and subscripts mod and era refer to quantities from the model and ERA5 respectively.

To assess the impact that a mean flow bias is having on the blocking statistics, we can replace $\overline{zg}_{mod}$ with $\overline{zg}_{era}$ in (8), recalculate the blocking statistics and assess the extent to which the bias goes away. This is referred to as "Mean fixed" in the figures. Conversely, to assess the impact of the variability bias, we can replace $zg'_{mod}(t)$ in (8) with $zg'_{era}(t)$ and recalculate the blocking statistics. This will be referred to as "Var fixed" in the figures.

There are many situations in which the interpretation of this decomposition will not be straightforward or quantitative. For example, the sum of the "Var fixed" and "Mean fixed" errors may be greater than the actual error due to compensating errors in both aspects. Furthermore, since the mean and the transients are intimately coupled, in a situation where both are in error, the decomposition gives no indication of the ultimate origins of the bias. For example, both the mean flow and transients may contribute, but it is possible that all that needs to be fixed is the transients and the mean flow will follow, or vice versa. Nevertheless, in situations where the contributions add up to the total error, this can provide a meaningful assessment of the relative influence of mean state biases versus biases in the representation of transient systems.

6.4.1 Northern Hemisphere

Systematic errors in the representation of blocking were first recognized in the numerical weather prediction context (Tibaldi & Molteni, 1990) and subsequently in early model intercomparisons (D'Andrea et al., 1998). Over the last two decades, substantial improvements in the modelled representation of Pacific blocking have occurred, but a reduced blocking frequency compared to reanalysis continues to be a pervasive issue in the North Atlantic, particularly during the winter (Vial & Osborn, 2012; Masato et al., 2013b; Anstey et al., 2013; Dunn-Sigouin & Son, 2013). Scaife et al. (2010) argued that this issue does not arise from the inability of models to capture the synoptic systems, but rather from biases in their mean state that prevent such systems from triggering the gradient reversal thresholds used in many blocking indices. The importance of mean state biases has now been identified in many studies for this region, but it is often not the only contributor (Vial & Osborn, 2012; Anstey et al., 2013; Davini & D'Andrea, 2016).

Fig. 15 presents blocking statistics for the NH. CESM2 captures the three dominant centers of blocking action during the winter: over Europe, Greenland and Eastern Siberia (compare Figs. 15a and b). In CESM1 (LENS), there was too much blocking over Eastern Siberia and Alaska (Fig. 15d) and this issue is now fixed in CESM2, but this improvement is due to an improvement in the mean state as opposed to any difference in the nature of synoptic anticyclone variability in this region (not shown). As is common with many models, including those from CMIP6 (Fig. A2e), both CESM1 and CESM2 have reduced blocking over the European and Greenland regions compared to observed. The Greenland blocking bias has been reduced in CESM2 and Gettelman et al. (2019) showed this improvement is larger in WACCM than in CAM (their Fig. 15). We see the same improvement in the Greenland blocking sector in WACCM with the blocking metric used here but larger biases occur in WACCM in the North Pacific (Fig. S17e) and so it does not appear as an improvement in our hemispheric error metric (Fig. 15i).

Figs. 17 a-c show that the BCAM6 NH winter biases can be roughly linearly decomposed into contributions that arise from the bias in the mean state and contributions that arise from the bias in the variability. This reveals that if the mean state bias is removed by artificially replacing CAM's mean state with ERA5's before calculating the blocking statistics, much of the European bias and roughly half of the Greenland bias goes away (Fig. 17b). Similar conclusions can be drawn for the European blocking bias in the CMIP6 ensemble mean (Fig. A2e and i). This indicates that a large fraction of the problem in this region lies in the mean state and the fact that it limits the potential for persistent anticyclones to reverse the zg gradient, as opposed to there being a problem with the variability itself. Nevertheless, the lack of variability is still contributing to roughly half of the Greenland bias and around a quarter of the bias off the coast of Europe in CESM2. Overall, in DJF, CESM2 has a very good representation of NH blocking statistics, is improved over CESM1 and is one of the top ranking models in this regard (Fig. 15i).

During the summer, CESM2 has unfortunately seen some degradations in the representation of blocking compared to LENS (Fig. 15g versus h). Too much blocking occurred in the highest latitudes in LENS and this issue remains in CESM2, but now this is accompanied by a severe lack of blocking in a latitude band to the south. This includes a lack of blocking in the observed centers of action over Russia, Iceland and Alaska. If this degradation were due to a change in the representation of the variability, then this would be of concern for the simulation of heat waves in these regions (Schaller et al., 2018). However, Figs. 17d-f make clear that this particular bias is associated with a mean state bias, as opposed to a variability bias and, indeed, it is the mean state change between LENS and CESM2 as opposed to a variability change that has given rise to this degradation. The zg gradient has strengthened around the Arctic circle, in association with the westerly bias discussed in section 4.2. This limits the ability of synoptic variability to induce a zg gradient reversal. A similar lack of summertime blocking over Northern Europe and western Russia is seen in the CMIP6 models (Fig. A2f) which can be partially attributed to a mean state bias and partially to a variability bias (Figs. A2j and n). Overall, the representation of summertime blocking in CESM2 has degraded, but this is largely a result of mean state changes and the errors due to the representation of the synoptic variability itself are relatively minimal (Fig. 17e).

6.4.2 Southern Hemisphere

During the wintertime (Fig. 16a), blocking occurs primarily in the Pacific sector (Sinclair, 1996; Berrisford et al., 2007), where the SH jet stream splits giving rise to blocking favourable conditions (Trenberth & Mo, 1985). During the summertime, blocking occurs primarily in the west Pacific/New Zealand sector but is less prevalent overall (Fig. 16e). In comparison to the NH, relatively little attention has been paid to the representation of blocking in the SH. Ummenhofer et al. (2013) assessed the representation of

blocking in the New Zealand sector of an older version of CAM (CAM3) and found the preferred blocking locations and seasonality to be well represented but with a systematic lack of blocking occurrence, which they related to the overly zonal flow in that region. Recently, Patterson et al. (2019) assessed the representation of SH blocking in the CMIP5 models and found that, while individual models can be substantially biased, there was a lack of consistency in the sense of the biases across models, except in the region south of Australia during summer where the majority of models exhibit a lack of blocking.

An assessment of SH blocking is provided in Fig. 16. The blocking index identifies substantial activity around the Antarctic continent, but this is in a region of relatively weak westerly winds and Berrisford et al. (2007) argue that it is debatable whether these should be considered blocking events. Therefore, we focus on mid-latitude blocking events by masking out the regions poleward of 60°S and evaluate the blocking characteristics between 30°S and 60°S.

During SH winter, the representation of blocking has unfortunately degraded in CESM2 compared to LENS. LENS only had a slight underestimation of blocking in the New Zealand sector (Fig. 16d), but CESM2 now also substantially underestimates blocking in the East Pacific center of action (Fig. 16c). This lack of blocking can be almost entirely ascribed to a degradation of the mean state (Fig. 17g-i) and is related to the fact that the westerlies have become too strong in the East Pacific, preventing anticyclonic circulations from trigerring the zg gradient reversal threshold (Fig. 3g). This is not an issue that is common to the CMIP6 models (Fig. A2g), but it seems that other CMIP6 models may have compensating errors that lead to a reasonable blocking climatology in this region, as Fig A2k shows that if the CMIP6 variability were placed on top of the ERA5 climatology, the models would consistently underestimate blocking in this region. Overall, the *NMSE* for JJA (Fig. 16i) should be viewed with caution for some models and for CESM2, since the SVR is considerably less than one indicating that the *NMSE* might be artificially reduced through the lack of spatial variance. Nevertheless, it reveals that it is more common for the CMIP6 models to overestimate blocking in the winter, rather than underestimate it like CESM2 does (the prevalence red circles in Fig. 16i). Some models indicate too much spatial variance while others exhibit too little and there are hardly any models where the SVR is close to one.

CESM2 also lacks SH blocking during the summer season in the West Pacific (Fig. 16g) which is rather different from the bias that was present in LENS (Fig. 16h). Again, this can be approximately linearly decomposed into contributions from errors in the mean state and errors in the variability (Figs. 17j-l). South of New Zealand, biases in the mean state dominate, where the anomalously strong westerlies and associated strong zg gradient, prevents anticyclones from overturning the zg gradient. In contrast, the deficiency in latitudes north of New Zealand is clearly dominated by a lack of variability in that region. The lack of blocking to the North of New Zealand is also common across the CMIP6 models (Fig. A2h), but biases in the mean state seem to play a more important role in this in CMIP6 (Figs. A2l and p). Again, the SVR in this season indicates that the *NMSE* should be viewed with caution as many models, including CESM, underestimate the spatial variance, while some others overestimate it (Fig. 16j).

7 Summary

A summary of CESM2's representation of many of the features discussed above, as compared to other CMIP models, is provided in Fig. 18. Here, all CMIP5 and CMIP6 models have been grouped together (excluding CESM2 versions). The ensemble mean LENS, coupled CESM2-CAM6 and coupled CESM2-WACCM6 have been given a ranking within this CMIP distribution based on either proximity to the observed value for metrics such as jet latitude, or the magnitude of the *NMSE* for spatial fields. For LENS,

a minimum to maximum range is provided based on the ranking of the poorest and highest performing of the 40 individual members. Here, we highlight some of the main conclusions of this analysis:

- **SH jet stream:** CESM2 ranks highly in SH jet position, as did CESM1. It is unusual in having the jet stream in the correct place, while the majority of CMIP5 and CMIP6 models place it too far equatorward. Degradations have occured in the representation of SH jet speed in CESM2 as the westerlies have become too strong in all seasons. This places CESM2 at the poorest end of the model distribution in the SH jet speed and 850hPa zonal wind assessments (Fig. 18b) and is an aspect that warrants particular attention in future development, given the global importance of southern ocean wind stress. Preliminary analysis indicates that parameter changes within the ice microphysics scheme are the primary contributor to this strengthening of the SH westerlies.
- **NH jet streams**: Substantial improvements in the representation of the NH wintertime 850hPa zonal wind are found in CESM2 compared to CESM1. The only biases of note that remain in this field are westerlies that are too strong over Europe and easterlies that are too strong over Africa - an issue that is very common among the CMIP6 models (Fig. A1e). The NH summertime jet streams have degraded slightly with a westerly bias that has developed around the Arctic circle leading to an Atlantic jet stream that is too fast and too poleward and a Pacific jet stream that is too fast.
- **Storm tracks:** A major advance in CESM2 over CESM1 is an improvement in the representation of storm tracks in both hemispheres and all seasons, as represented by 850hPa 10-day high pass filtered eddy meridional wind variance (Figs. 3, 6, S2 and S5). LENS and many other CMIP models exhibit a hemispheric lack of storm track activity (an unconditional bias) in both the NH and SH. In CESM2 this has been alleviated, leaving only smaller remaining phase errors, making CESM2 a high ranking model in this aspect. A particularly notable improvement is found in the lee of the Andes and the Rockies. Here, the real world exhibits substantial lower tropospheric meridional wind variance that was almost entirely absent in LENS, but is now represented with great fidelity in CESM2. Most of this improvement has arisen from the changes in the representation of turbulent orographic form drag and, to a lesser extent, mesoscale orographic blocking (see Section 4.3).
- **Stationary waves**: Given the importance of stationary waves for the representation of regional climate, it is important that they be simulated with accuracy. CESM2 is one of the highest ranking models in the representation of NH stationary waves in both winter and summer (Fig. 18a) and is substantially improved compared to CESM1. This comes with the additional caveat that compensating errors contribute to the summertime representation, as it degrades when observed SSTs are prescribed (see section 5.3). In the SH, along with the degradation of the zonal mean state, the representation of stationary waves has degraded in CESM2 compared to CESM1, but it still ranks at roughly the middle of the CMIP range (Fig. 18b).
- **Divergent circulation**: The divergent circulation is closely connected to tropical precipitation and represents an important forcing of extra-tropical stationary waves. CESM2 has a remarkable representation of the upper tropospheric velocity potential in both summer and winter (Fig. 18c) with an *NMSE* that is almost as small as the difference between ERA5 and ERA-Interim reanalysis during summer. Again, for the summertime, this does come with the caveat that its representation is degraded when observed SSTs are prescribed, pointing to compensating errors. Nevertheless, this is a field that has been substantially improved compared to CESM1. Any changes in Hadley cell metrics (Fig. 18c, right) should be viewed with caution given the lack of consistency across reanalysis datasets, but overall, CESM2 lies close to the reanalysis range.

- **SAM:** The SAM is the dominant mode of variability in the extra-tropical SH circulation and future projected circulation change is expected to project strongly onto the SAM. Thus, the SAM should be simulated accurately in order to represent both variability and future climate change with fidelity. The zonal mean structure of the daily zg SAM and the SAM persistence are represented well in CESM2. The latitude-longitude structure of zg anomalies associated with the SAM has seen some degradation with potential consequences for surface temperature and sea ice variability around the Antarctic peninsula. Despite this degradation, CESM2 still ranks highly in terms of SAM structure (Fig. 18b, right).

- **NAM:** Much like the SAM, the NAM is the dominant mode of extra-tropical zonal mean circulation variability in the NH and future predicted circulation changes are expected to project strongly onto the NAM. The NAM structure and persistence are also well represented in CESM2 with slight improvements compared to CESM1. A ranking is not provided for NAM persistence in Fig. 18a because of the large sampling uncertainty that is present for this quantity.

- **NAO:** The NAO is closely connected to the NAM but is more localized in the North Atlantic sector. Accurate simulation of the NAO is important for the representation of climate variability over much of the North Atlantic sector. CESM2 represents the structure of the winter and summer NAO well, but the large sampling uncertainty on individual members and in the observations means it is challenging to be quantitative about this.

- **Blocking:** Blocking is relevant for the simulation of extreme weather events in the mid-latitudes. CESM2 is one of the highest ranking models in terms of NH blocking (Fig. 18a). While there is still an underestimation of blocking frequency in the Greenland and European sectors in winter, a large fraction of this can be ascribed to errors in the mean state that prevent persistent anticyclones from overturning the zg gradient, as opposed to an error in the variability itself. While summertime blocking has degraded in CESM2 compared to CESM1, this is primarily due to a degradation of the mean state as opposed to a change in the nature of the synoptic variability. We choose not to rank SH blocking in this figure given the issues that arise with the *NMSE* calculation resulting from the widely varying representation of spatial variance across the models. SH blocking is poorly represented in CESM2 but again, this is primarily because of biases in the mean state zg gradient as opposed to the variability being in error, except for in the region north and east of New Zealand during DJF.

The majority of the features assessed here were not considered directly during the tuning processes with only some attention paid to the representation of the climatological *slp*, wind stress and precipitation fields. Many of the improvements (and degradations) are, therefore, emergent properties of this new model, presumably arising from the upgraded physical parameterizations or improved tuning of basic aspects of the mean climate such as precipitation and *slp*. Overall, despite some degradations, CESM2 exhibits many improvments over CESM1. It is a high ranking model in most aspects of the atmospheric circulation and will be a useful tool for the study of many aspects of climate variability and change.

Table 1. Summary of simulations used. The top portion of the left three columns depict the model number (used to depict the model in each plot), model name and number of members of each CMIP5 model. The top portion of the right hand columns show the same for CMIP6. A "*" at the end of the CMIP model name depicts whether that model is used in the analyses requiring daily *ua* or *va* and a "+" depicts whether a model is used in analysis requiring daily *zg*. The lower portion of the left columns summarizes the CESM1 and CESM2 simulations. The period from 1979 to 2014 is used for all simulations.

CMIP5				CMIP6		
#	Name	Members		#	Name	Members
1	ACCESS1-0	1		1	ACCESS-CM2*+	2
2	ACCESS1-3	1		2	ACCESS-ESM1-5*	3
3	bcc-csm1-1	1		3	AWI-CM-1-1-MR	5
4	bcc-csm1-1-m	1		4	BCC-CSM2-MR*+	3
5	BNU-ESM*+	1		5	BCC-ESM1*+	3
6	CanESM2*+	5		6	CAMS-CSM2-0	1
7	CCSM4*+	6		7	CanESM5*+	25
8	CESM1-CAM5	3		8	CNRM-CM6-1*+	15
9	CESM1-WACCM	1		9	CNRM-CM6-1-HR*+	1
10	CMCC-CM	1		10	CNRM-ESM2-1*+	5
11	CMCC-CMS	1		11	E3SM-1-0	5
12	CNRM-CM5*+	5		12	E3SM-1-1	1
13	CSIRO-Mk3-6-0*	10		13	EC-Earth3*+	10
14	FGOALS-g2	1		14	EC-Earth3-Veg*	4
15	FIO-ESM	1		15	FGOALS-f3-L	3
16	GFDL-CM3*+	1		16	FGOALS-g3	3
17	GFDL-ESM2G*+	1		17	FIO-ESM-2-0	3
18	GFDL-ESM2M*+	1		18	GFDL-CM4*+	1
19	GISS-E2-H	1		19	GFDL-ESM4	1
20	GISS-E2-R	1		20	GISS-E2-1-G*+	10
21	HadGEM2-AO	1		21	GISS-E2-1-G-CC	1
22	HadGEM2-CC	1		22	GISS-E2-1-H	10
23	HadGEM2-ES	1		23	HadGEM3-GC31-LL*+	4
24	inmcm4	1		24	HadGEM3-GC31-MM*+	2
25	IPSL-CM5A-LR*+	4		25	INM-CM4-8*	1
26	IPSL-CM5A-MR	1		26	INM-CM5-0*	10
27	IPSL-CM5B-LR	1		27	IPSL-CM6A-LR*+	32
28	MIROC5*+	3		28	KACE-1-0-G	3
29	MIROC-ESM*+	1		29	MCM-UA-1-0	1
30	MIROC-ESM-CHEM*+	1		30	MIROC6*+	10
31	MPI-ESM-LR*+	3		31	MIROC-ES2L*	3
32	MPI-ESM-MR*+	1		32	MPI-ESM-1-2-HAM*+	2
33	MRI-CGCM3*+	1		33	MPI-ESM1-2-HR*+	10
34	NorESM1-M*+	1		34	MPI-ESM1-2-LR*+	10
35	NorESM1-ME	1		35	MRI-ESM2-0*+	5
CESM1				36	NESM3	5
	LENS	40		37	NorCPM1	30
CESM2				38	NorESM2-LM*+	3
	BCAM6	11		39	NorESM2-MM*	1
	BWACCM6	3		40	SAM0-UNICON*	1
	FCAM6	3		41	TaiESM1	1
	FWACCM6	3		42	UKESM1-0-LL*+	4

Table 2. Summary of CESM2 simulations. From left to right: simulation name; approximate horizontal resolution in degrees longitude×latitude format; number of vertical (mid-) levels and the model lid pressure; and a description of external forcings and the form of the SSTs or parameterization changes. For FCAM6*, FCAM5, FCAM6_TMS and FCAM6_NOMOB, we use 1979-2005, while for the other simulations we use 1979-2014.

Name	Res (lon×lat)	# levels/lid p	Description
BCAM6	1.9°×2.5°	32/2.26hPa	historical, coupled ocean
BWACCM6	1.9°×2.5°	70/4.5e-6hPa	historical, coupled ocean
FCAM6	1.9°×2.5°	32/2.26hPa	historical, prescribed observed SSTs (Hurrell et al., 2008)
FWACCM6	1.9°×2.5°	70/4.5e-6hPa	historical, prescribed observed SSTs (Hurrell et al., 2008)
FCAM6MOD	1.9°×2.5°	32/2.26hPa	historical, prescribed SSTs from member 11 of BCAM6
FCAM6*	1.9°×2.5°	32/2.26hPa	as FCAM6 but with land biogeochemistry turned off
FCAM5	1.9°×2.5°	32/2.26hPa	as FCAM6* but with CAM5 physics
FCAM6_TMS	1.9°×2.5°	32/2.26hPa	as FCAM6* but with the Beljaars scheme replaced by TMS
FCAM6_NOMOB	1.9°×2.5°	32/2.26hPa	as FCAM6* but with mesoscale orographic blocking turned off
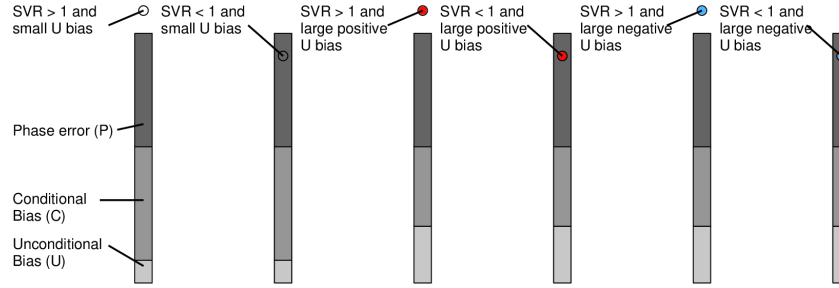
Figure 1. An explanation of the representation of the NMSE, SVR and unconditional bias in each figure. The NMSE is depicted by vertical bars with three shaded components (where relevant). The three components are, from bottom/light to top/dark, the Unconditional Bias, Conditional Bias and the Phase Error (see Equation (3)). The SVR is depicted by the circle. Where the SVR lies above the bar as in the 1st, 3rd and 5th bars here, the SVR is greater than 1 and vice-versa. When the magnitude of the spatial mean bias is more than 10% of the ERA5 spatial mean, we consider that to be a "large" unconditional bias and shade the SVR circle red for positive (3rd and 4th bar examples) and blue for negative (5th and 6th bar examples) biases. Otherwise, the SVR circlce is left open.
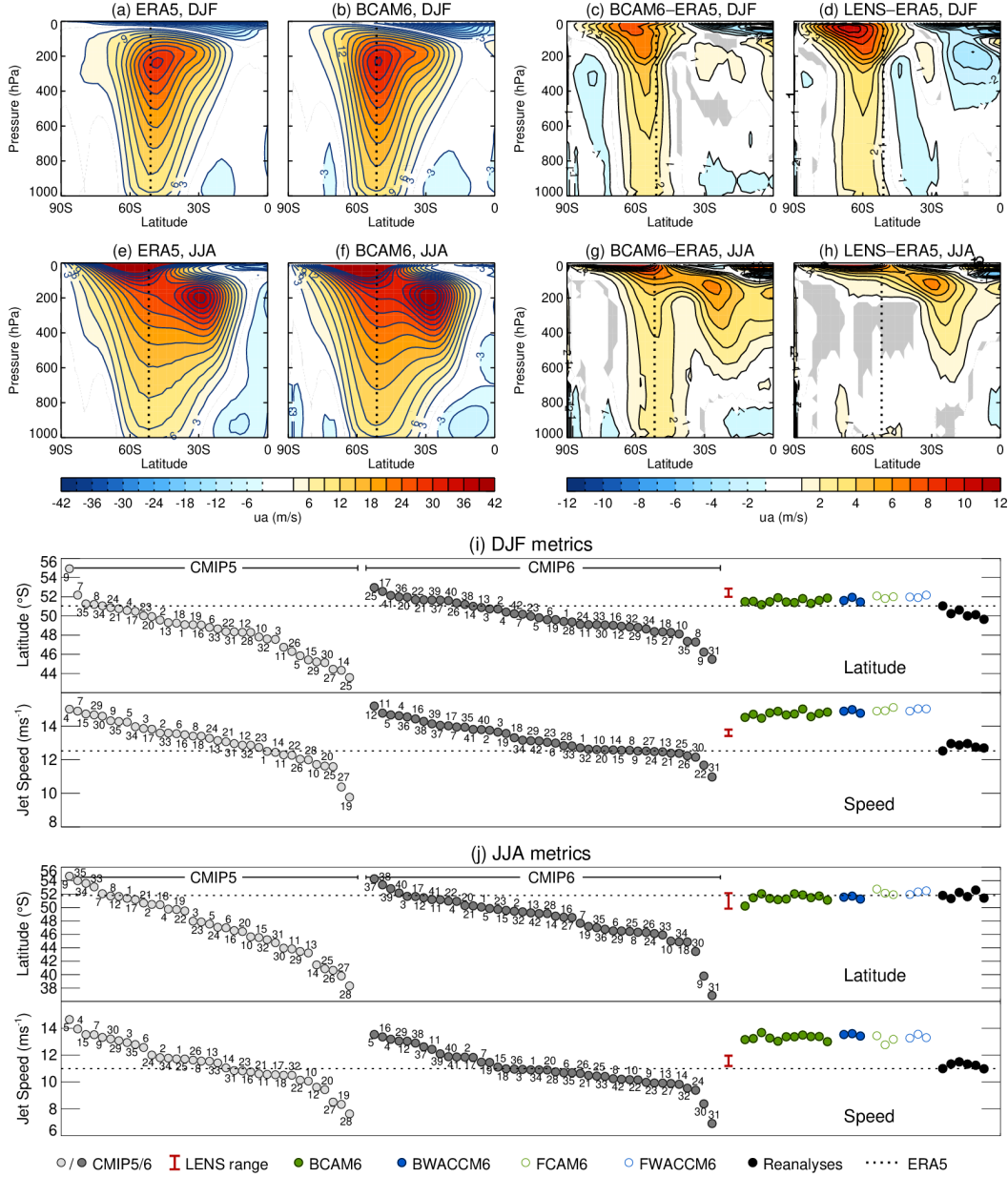
Figure 2. (a)-(d) DJF zonal mean zonal wind for (a) ERA5, (b) BCAM6, (c) BCAM6−ERA5, (d) LENS−ERA5. The black vertical line in (a), (c) and (d) depicts the ERA5 jet latitude, while that in (b) depicts the BCAM6 jet latitude. Gray shading in (c) and (d) shows where ERA5 lies within the spread of the model members. (e)-(h) as (a)-(d) but for JJA. (i) shows the jet latitude (top) and jet speed (bottom) metrics for DJF. Light gray = CMIP5, dark gray = CMIP6, the light red range shows the range across the 40 LENS members, solid green = BCAM6, solid blue = BWACCM6, open green = FCAM6, open blue = FWACCM6 and black = reanalyses from left to right: ERA5, ERA-Interim, MERRA2, JRA55, ERA20C and 20CR. The horizontal dotted line shows the ERA5 values. (j) is as (i) but for JJA.
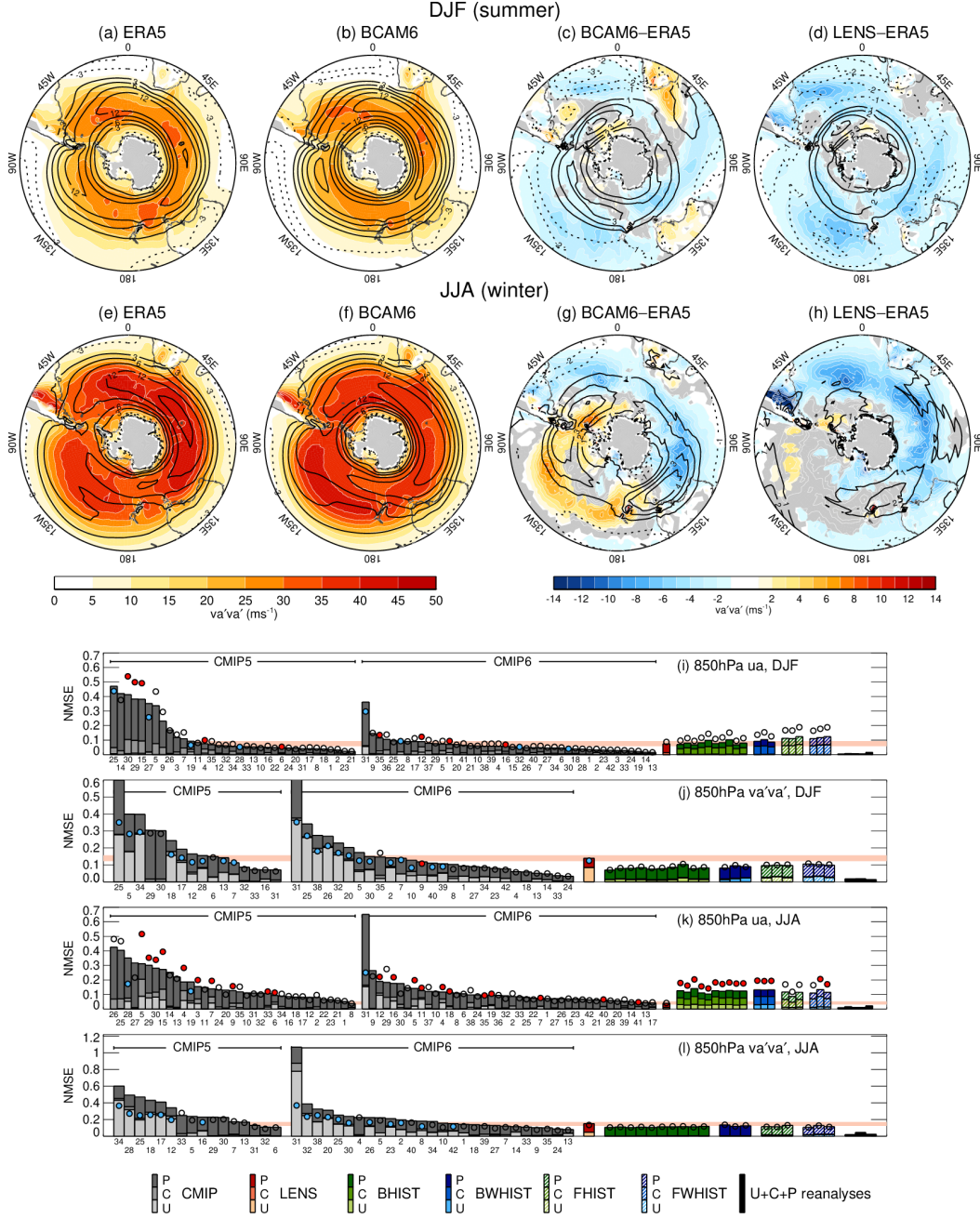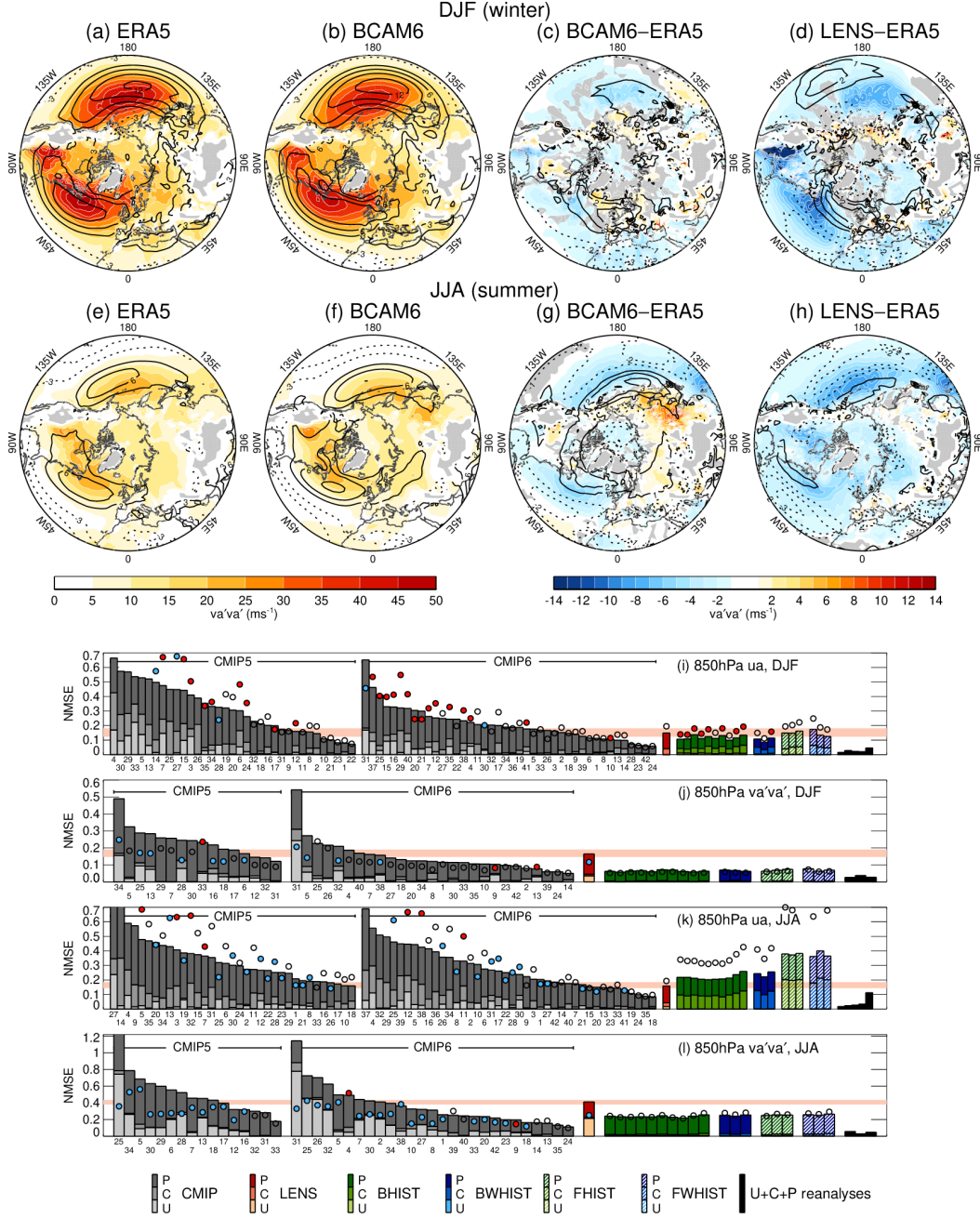
Figure 3. (a)-(h) 850hPa *ua* (contours) and 850hPa 10-day high pass filtered eddy meridional wind variance ($va'va'$) (shading), between 20°S and the pole. (a) ERA5, (b) BCAM6, (c) BCAM6−ERA5 and (d) LENS−ERA5 for DJF. (e)-(h) are as (a)-(d) but for JJA. Gray shading in (c), (d), (g) and (h) depicts where ERA5 $va'va'$ lies within the model ensemble spread. (i) and (j) show *NMSE* of DJF 850hPa *ua* and $va'va'$, respectively and (k) and (l) show the same for JJA. Bars depict the U, C and P contributions to the *NMSE* (Eq. 3) as described in the legend and circles depict the SVR. Circles are shaded red/blue when the unconditional bias corresponds to the magnitude of the mean bias being greater than 10% of the ERA5 spatial mean. Reanalyses are from left to right; ERA-Interim, MERRA2, JRA55 (and ERA20C and 20CR for *ua*). The light red range that extends across each panel (i)-(l) depicts the minimum to maximum range of *NMSE* from the LENS members.
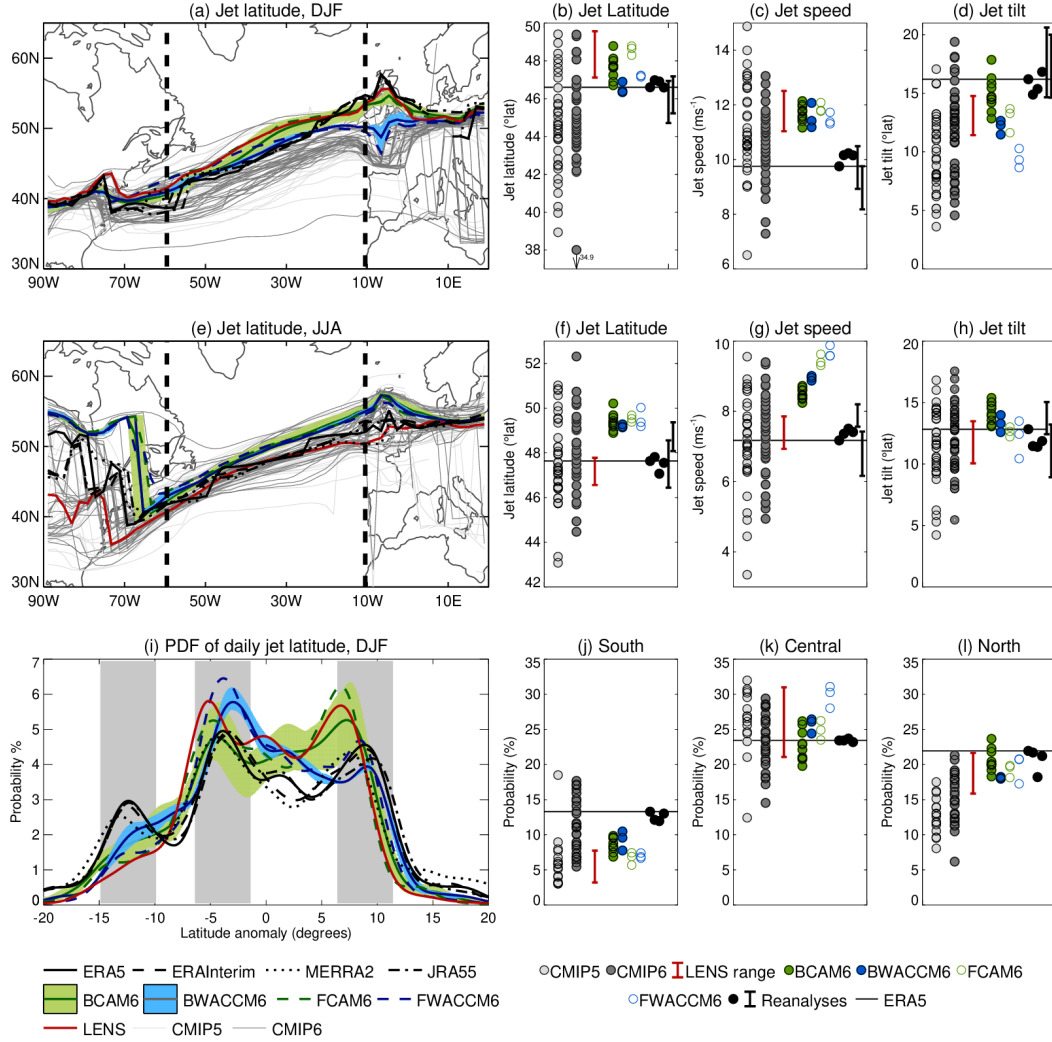
Figure 4. Metrics of the North Atlantic jet. (a) Local jet latitude, (b) Atlantic averaged jet latitude, (c) Atlantic averaged jet speed and (d) jet tilt across the Atlantic, during DJF. Note that one CMIP6 model is off the scale in (b), as indicated. Metrics are calculated between 60°W and 10°W (black dashed lines in (a)). (e)-(h) are as (a)-(d) but for JJA. Reanalyses are from left to right: ERA5, ERA-Interim, MERRA2, JRA55, ERA20C and 20CR. For ERA20C and 20CR we show the minimum to maximum range of all (overlapping) 36 year segments of the record. Values of these metrics for individual models are listed in Table S1. (i) the PDF of daily jet latitude with gray vertical shaded regions depicting the latitude range used for the probabilities shown in (j)-(l). (j)-(l) the probability of the jet-latitude being in the Southern, Central and Northern regions (gray regions in (i)). Green and Blue shading in (a), (e) and (i) depict the minimum to maximum range for the 11 BCAM6 and 3 BWACCM6 ensemble members, respectively. ERA20C and 20CR are not shown for the daily metrics.
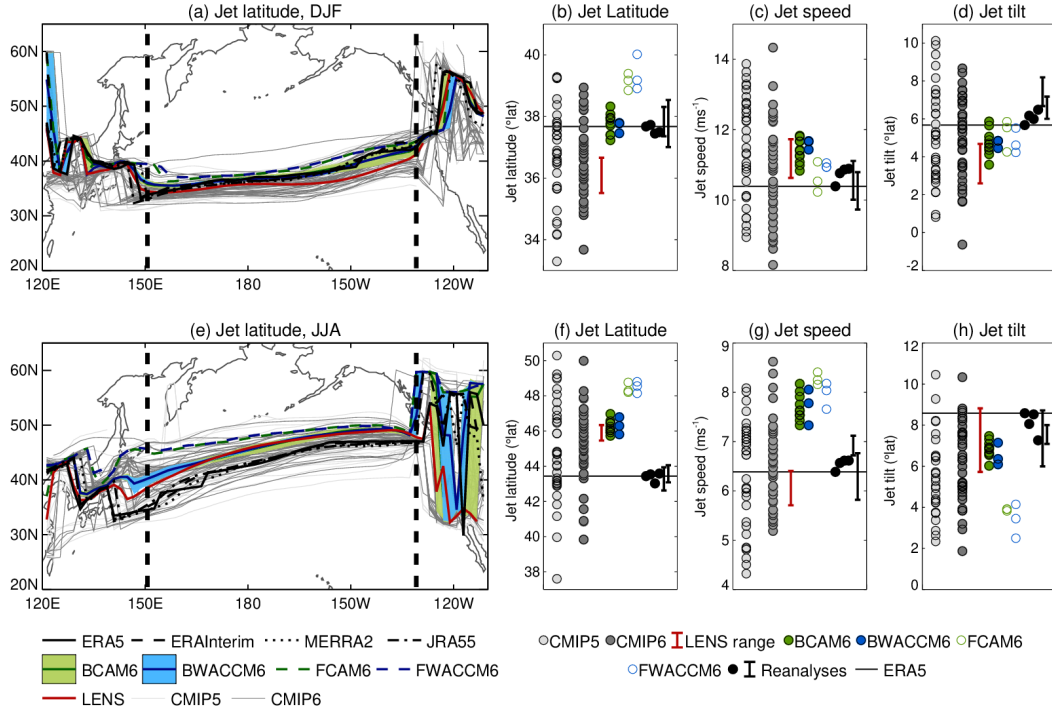
Figure 5. Metrics of the North Pacific jet. (a) Local jet latitude, (b) Pacific averaged jet latitude, (c) Pacific averaged jet speed and (d) the jet tilt across the Pacific, during DJF. Metrics are calculated between 150°E and 130°W (black dashed lines in (a)). Reanalyses are from left to right: ERA5, ERA-Interim, MERRA2, JRA55, ERA20C and 20CR. For ERA20C and 20CR we show the minimum to maximum range of all (overlapping) 36 year segments of the record. Values of these metrics for individual models are listed in Table S2. (e)-(h) are as (a)-(d) but for JJA. Green and blue shaded ranges in (a) and (e) depict the minimum to maximum range for the 11 BCAM6 and 3 BWACCM6 members.

Figure 6. (a)-(h) 850hPa *ua* (contours) and 850hPa 10-day high pass filtered eddy meridional wind variance ($va'va'$) (shading), between 20°N and the pole. (a) ERA5, (b) BCAM6, (c) BCAM6−ERA5 and (d) LENS−ERA5 for DJF. (e)-(h) are as (a)-(d) but for JJA. Gray shading in (c), (d), (g) and (h) depicts where ERA5 $va'va'$ lies within the model ensemble spread. (i) and (j) show *NMSE* of DJF 850hPa *ua* and $va'va'$, respectively and (k) and (l) show the same for JJA. Bars depict the U, C and P contributions to the *NMSE* (Eq. 3) as described in the legend and circles depict the SVR. Circles are shaded red/blue when the unconditional bias corresponds to the magnitude of the mean bias being greater than 10% of the ERA5 spatial mean. Reanalyses are from left to right; ERA-Interim, MERRA2, JRA55 (and ERA20C and 20CR for *ua*). The light red range that stretches across each panel (i)-(l) shows the minimum to maximum range of *NMSE* from the LENS members.
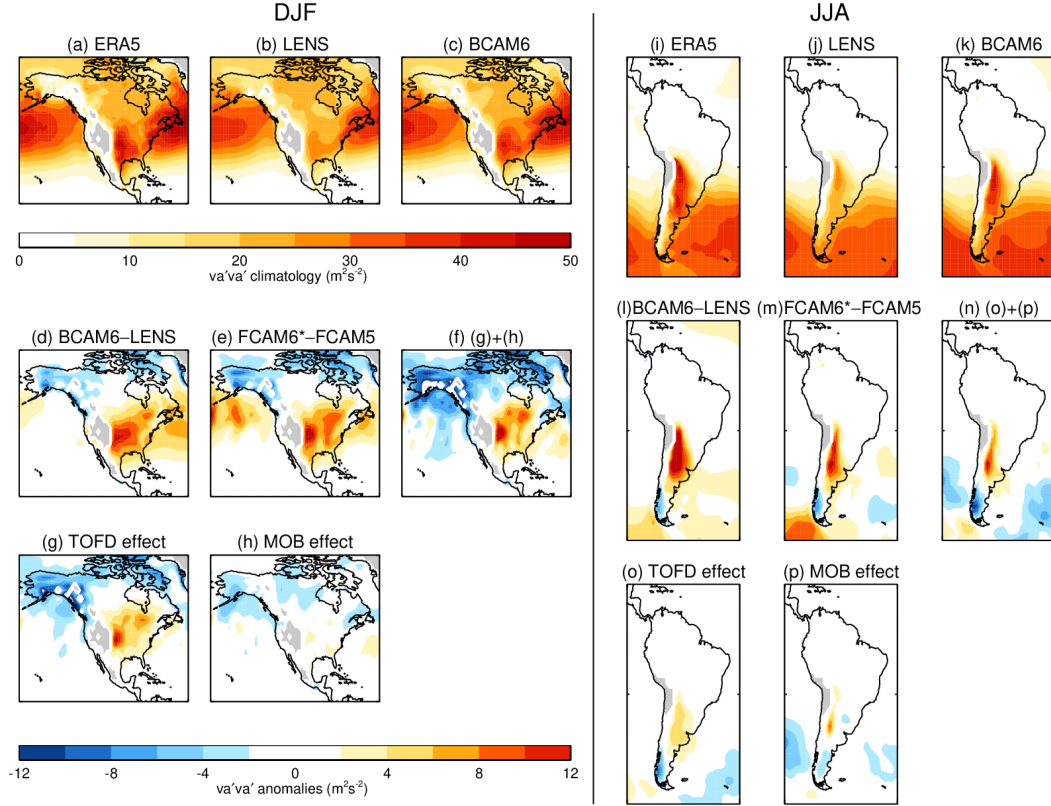
Figure 7. 10-day high pass filtered meridional wind variance $\overline{v_a'v_a'}$ in the lee of the Rockies (left) and the Andes (right) during winter. For the Rockies, (a)-(c) show the climatologies for ERA5, LENS and BCAM6. (d) shows BCAM6−LENS which can be compared with panel (e) which shows the FCAM6*−FCAM5 difference that forms the baseline for the sensitivity tests. (f) shows the sum of the differences FCAM6*−FCAM6_TMS and FCAM6*−FCAM6_NOMOB i.e., the influences of the TOFD and MOB schemes inferred by reverting them back to CAM5 settings. The individual differences FCAM6*−FCAM6_TMS and FCAM6*−FCAM6_NOMOB can be seen in panels (g) and (h). Panels (i-p) are as (a-h) but over South America during JJA. Gray regions are where the 850hPa level lies below the surface.
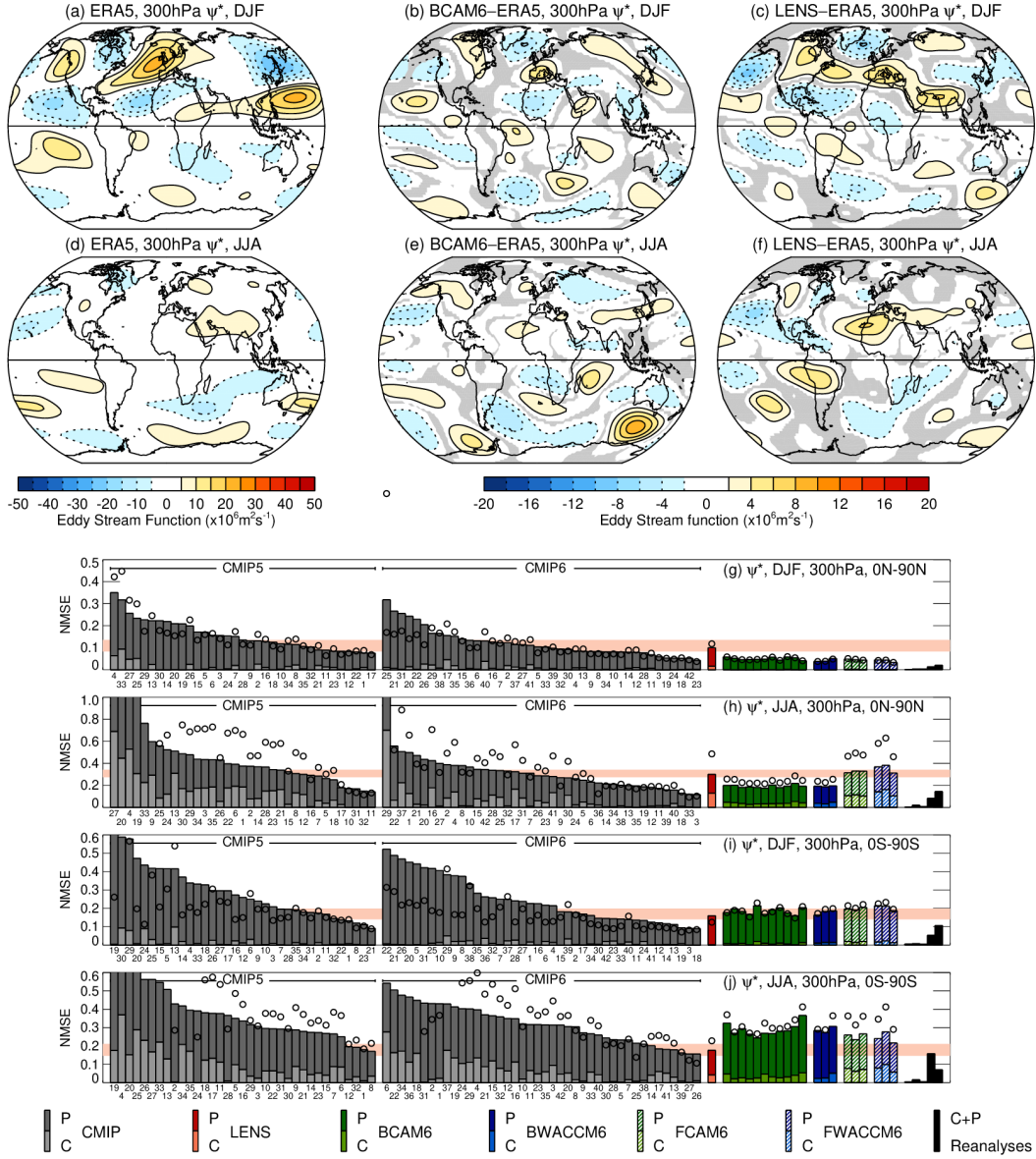
Figure 8. The representation of 300hPa eddy stream function (ψ^*). (a) ERA5, (b) BCAM6−ERA5 and (c) LENS−ERA5. Gray shading in (b) and (c) depicts where ERA5 lies within the spread of the model ensemble members. (d)-(f) are as (a)-(c) but for JJA. (g)-(j) shows the *NMSE* and its components (see legend) along with the SVR (circles) for NH DJF, NH JJA, SH DJF and SH JJA, respectively. The light red range that spans each panel (g)-(j) depicts the minimum to maximum range of *NMSE* of the 40 LENS members and the ordering of the reanalyses from left to right is ERA-Interim, MERRA2, JRA55, ERA20C and 20CR. Note there is no unconditional bias for ψ^* since its average around the longitude circle is zero by definition.
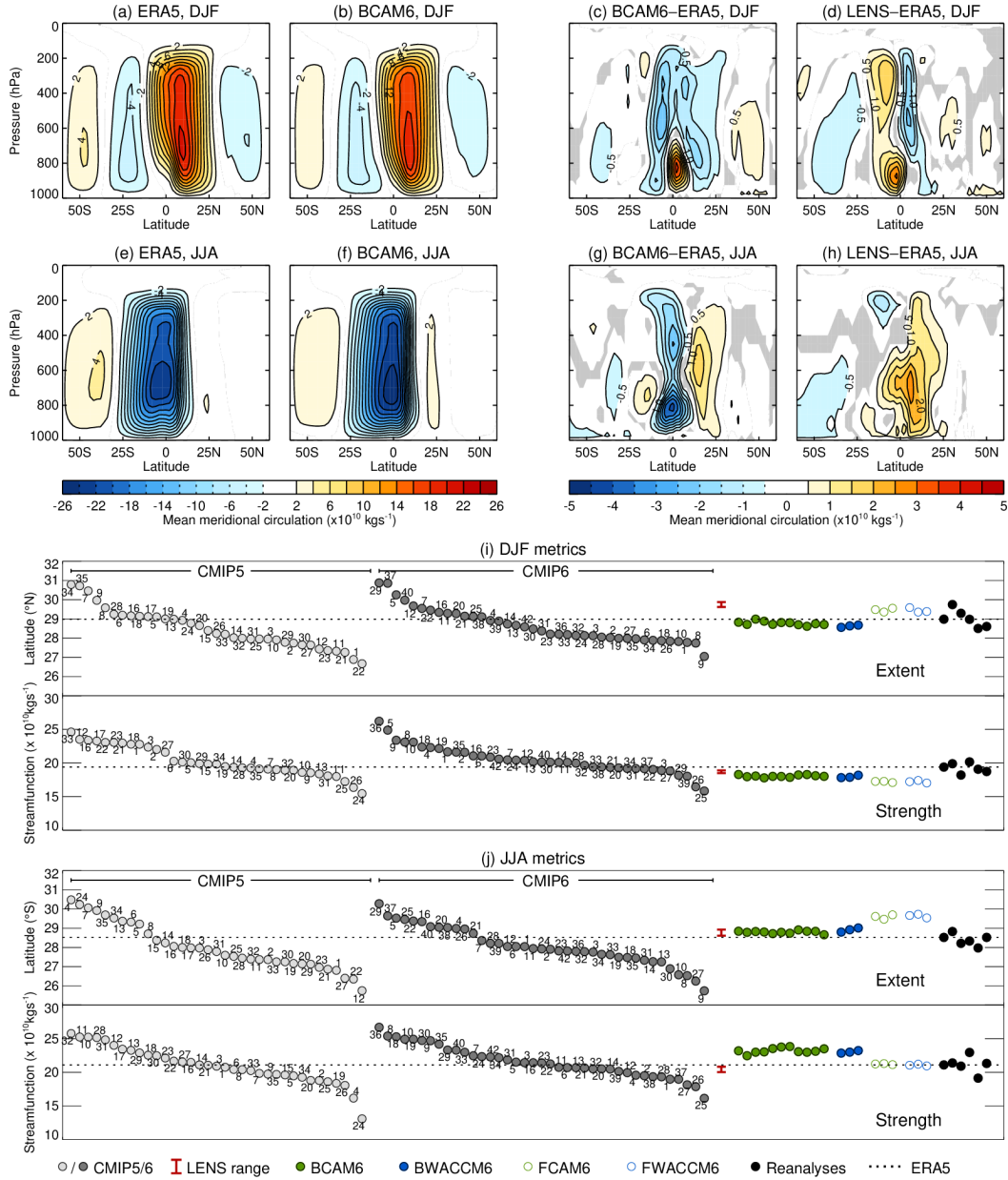
Figure 9. (a)-(d) The DJF zonal mean meridional stream function (Eq. 5) for (a) ERA5, (b) BCAM6, (c) BCAM6−ERA5 and (d) LENS−ERA5. Gray shading in (c) and (d) shows regions where ERA5 lies within the distribution of the model ensemble members. (e)-(h) are as (a)-(d) but for JJA. (i) DJF Hadley cell metrics: top = extent; bottom = strength . (j) as (i) but for JJA. The reanalyses are, from left to right: ERA5, ERA-Interim, MERRA2, JRA55, ERA20C and 20CR.
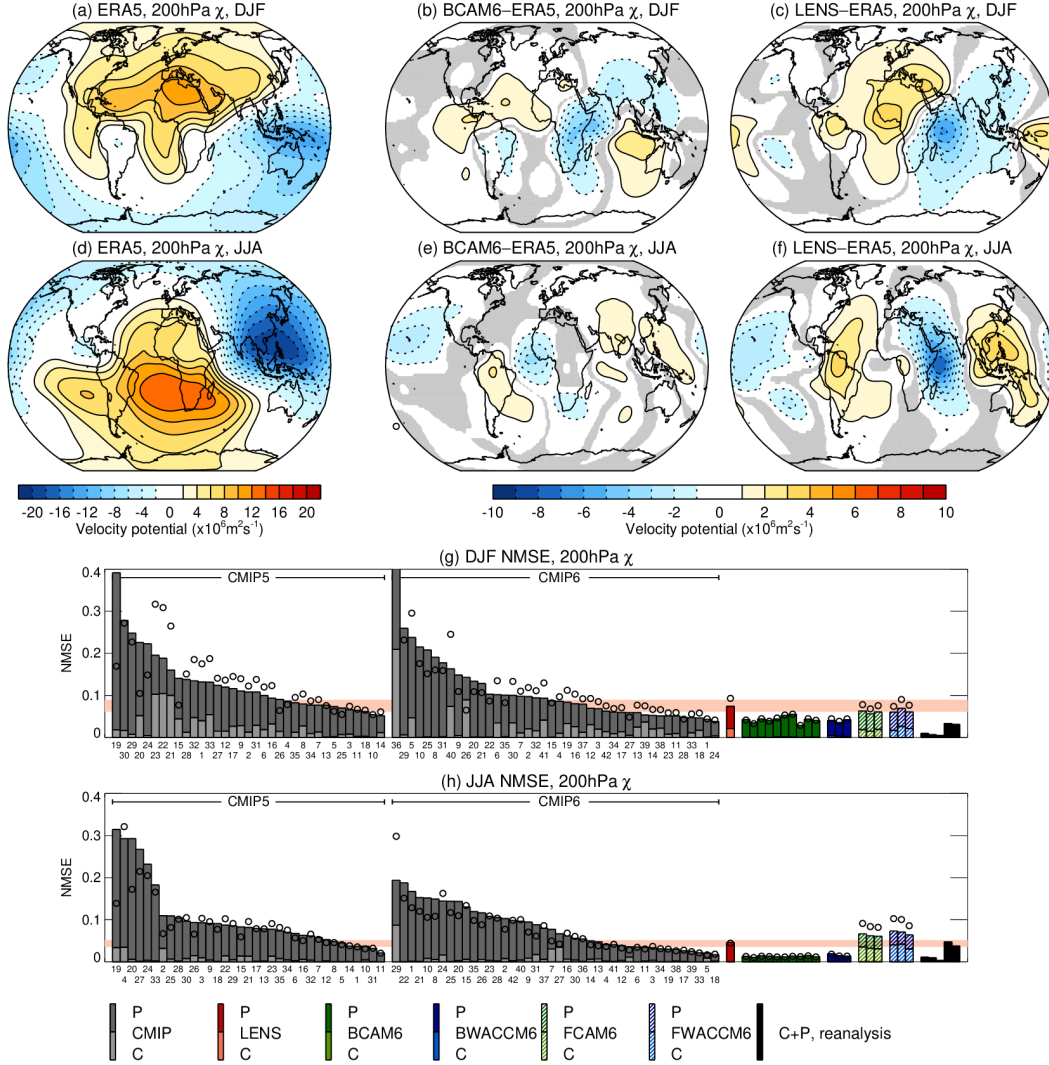
Figure 10. The representation of 200hPa velocity potential χ. (a) ERA5, (b) BCAM6−ERA5 and (c) LENS−ERA5 for DJF. Gray shading in (b) and (c) depicts where ERA5 lies within the spread of the model ensemble members. (d)-(f) are as (a)-(c) but for JJA. (g) and (h) show the global *NMSE* and its components (see legend) along with the SVR (circles) for DJF and JJA, respectively. The light red range that extends across panels (g) and (h) depicts the minimum to maximum range of the LENS ensemble members and the ordering of the reanalyses from left to right is ERA-Interim, MERRA2, JRA55, ERA20C and 20CR. Note there is no unconditional bias for χ since its global average is zero.
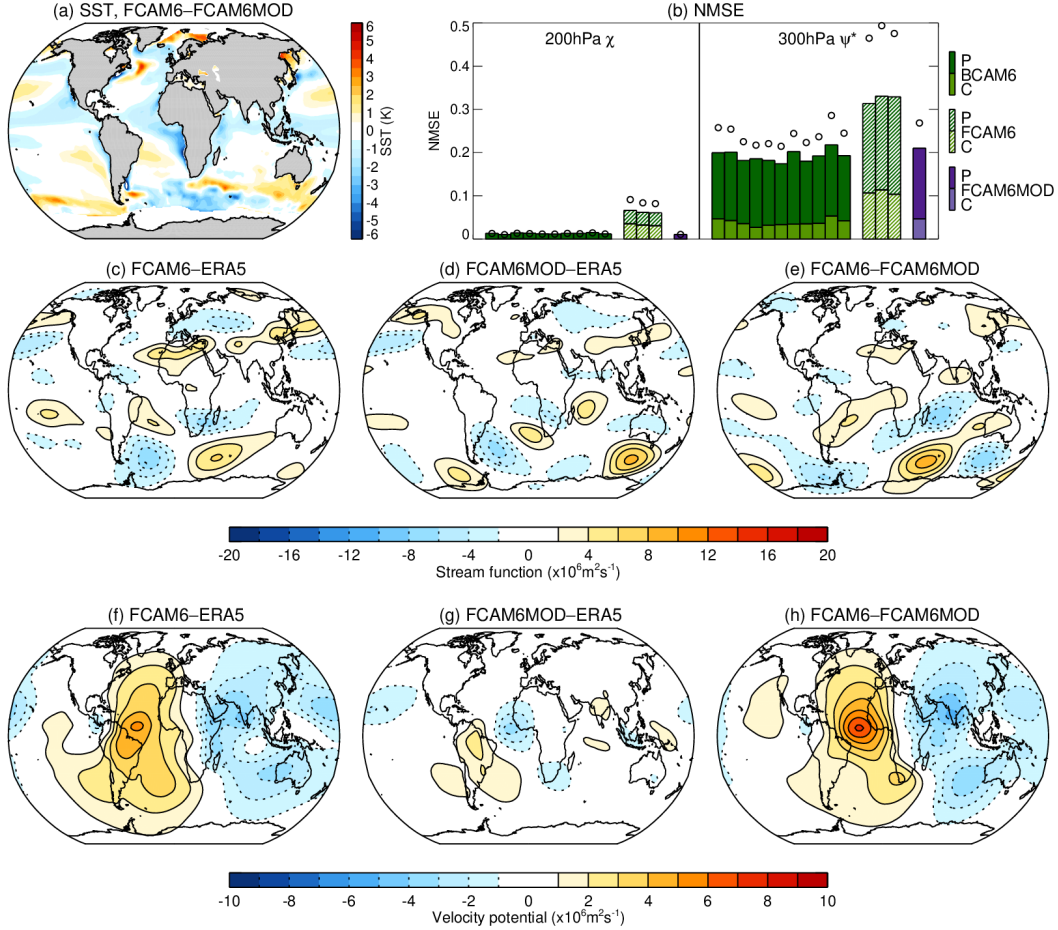
Figure 11. (a) The difference between the JJA climatology of the observed SSTs prescribed in FCAM6 and that of the modelled SSTs from the BCAM6 member used to provide SSTs for FCAM6MOD. (b) the JJA *NMSE* and SVR for (left) global 200hPa velocity potential and (right) 300hPa NH eddy stream function for BCAM6 and FCAM6 (repeated from Figs. 8h and 10h) and for FCAM6MOD (purple). (c)-(e) 300hPa eddy stream function for FCAM6−ERA5, FCAM6MOD−ERA5 and FCAM6−FCAM6MOD, respectively. (f)-(h) are as (c)-(e) but for 200hPa velocity potential.
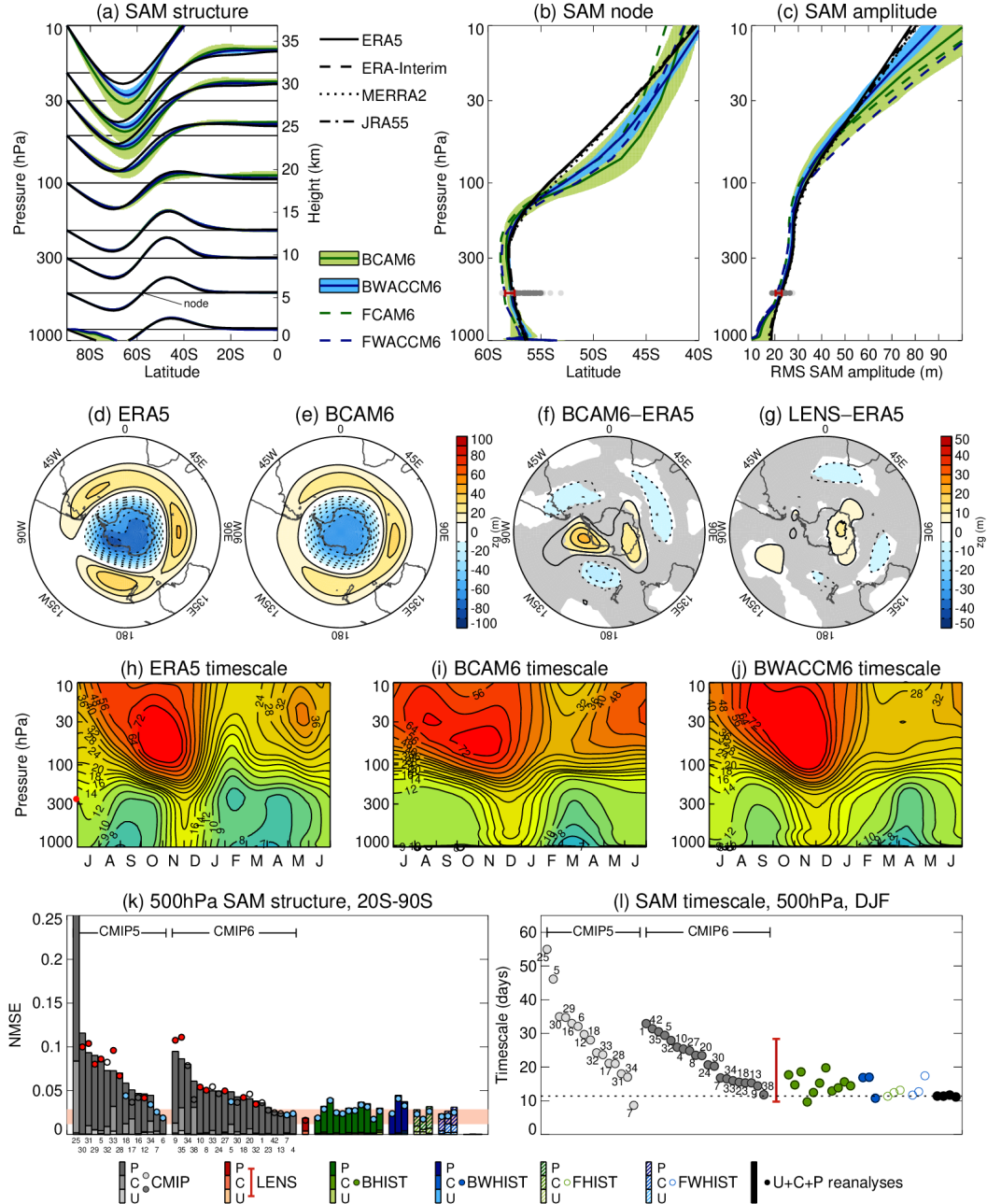
Figure 12. The representation of the SAM. (a) the SAM structure obtained by regressing zonal mean zg, weighted by cosine of latitude, onto the SAM, displacing by the height of the relevant pressure level and multiplying by 100 so that it can be viewed with the height axis of the panel. (b) The SAM node: the latitude of the zero crossing between the negative and positive nodes of the SAM (see 500hPa structure in (a)). (c) The root mean square amplitude of the SAM structure. Green and blue shading in (a)-(c) depicts the minimum to maximum range for the BCAM6 and BWACCM6 members. (d)-(g) the 500hPa zg structure of the SAM for ERA5, BCAM6, BCAM6−ERA5 and LENS−ERA5, respectively. Gray shading in (f) and (g) is where ERA5 lies within the model ensemble spread. (h)-(j) SAM timescale as a function of season and height for ERA5, BCAM6 and BWACCM6, respectively. (k) *NMSE* (see legend) and SVR (circles) of the 500hPa latitude-longitude SAM structure. The light red range depicts the minimum to maximum range of the LENS members. Reanalyses (barely visible) are, from left to right, ERA-Interim, MERRA2 and JRA55. (l) DJF averaged SAM timescales at 500hPa. Reanalyses are from left to right: ERA5, ERA-Interim, MERRA2 and JRA55 and black dashed line = ERA5.
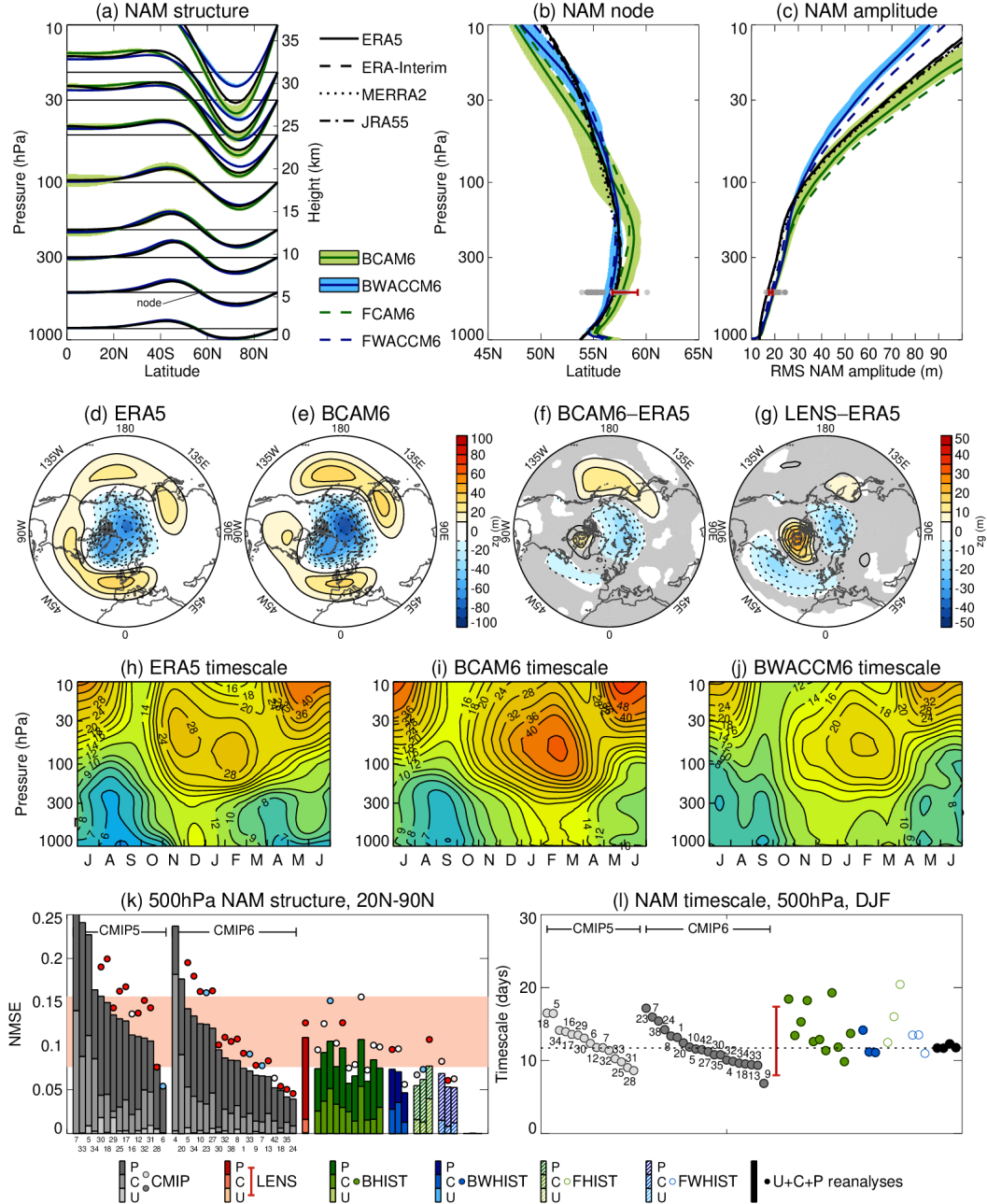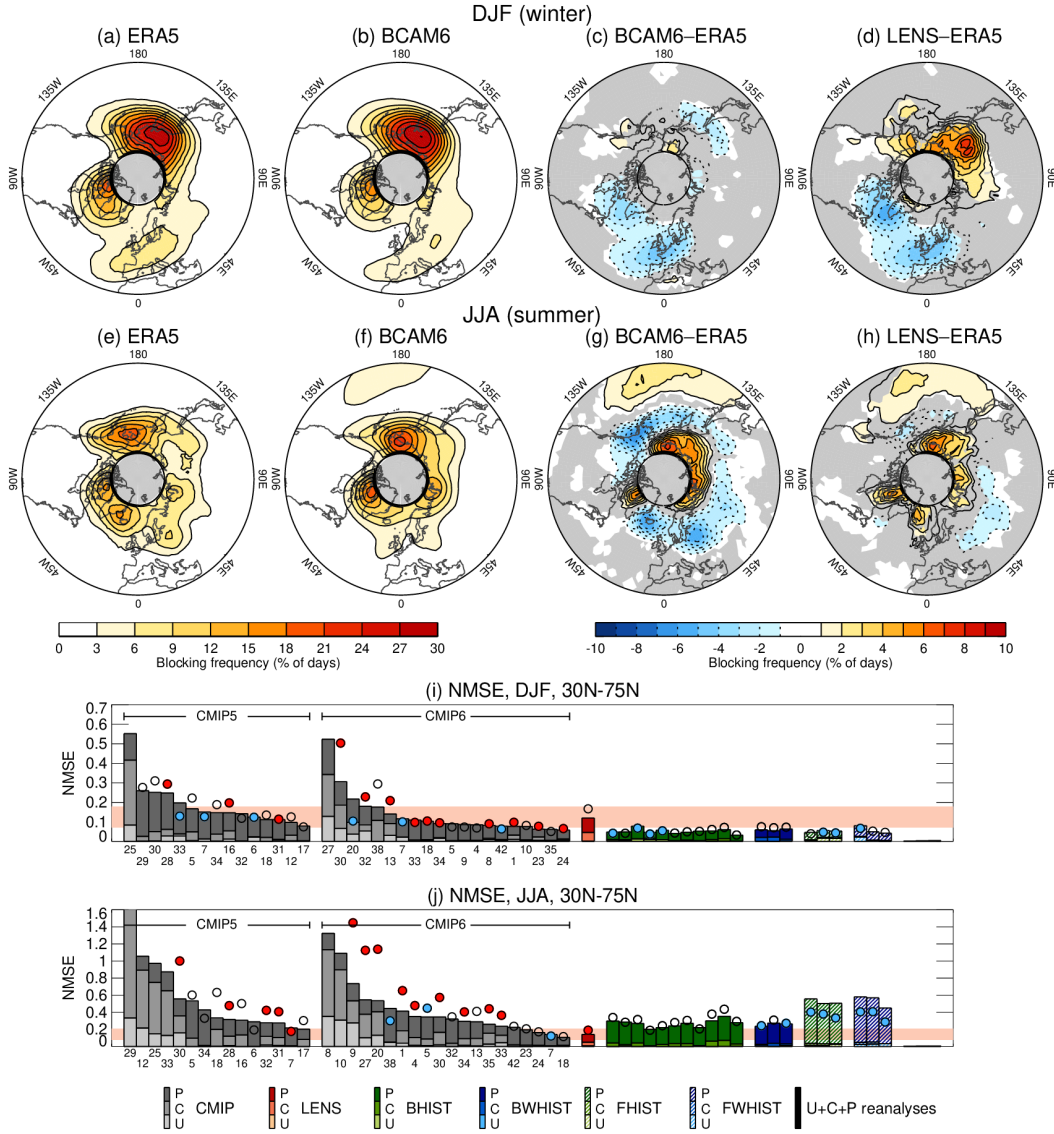
Figure 13. The representation of the NAM (a) the NAM structure obtained by regressing zonal mean zg weighted by cosine of latitude, onto the NAM, displacing by the height of the relevant pressure level and multiplying by 100 so it can be viewed with the height axis of the panel. (b) The NAM node: the latitude of the zero crossing between the negative and positive NAM nodes (see 500hPa structure in (a)). (c) The root mean square amplitude of the NAM structure. Green and blue shading in (a)-(c) depicts the minimum to maximum range for the BCAM6 and BWACCM6 members. (d)-(g) the 500hPa zg structure of the NAM for ERA5, BCAM6, BCAM6−ERA5 and LENS−ERA5, respectively. Gray shading in (f) and (g) is where ERA5 lies within the model ensemble spread. (h)-(j) NAM timescale as a function of season and height for ERA5, BCAM6 and BWACCM6, respectively. (k) *NMSE* (see legend) and SVR (circles) of the 500hPa latitude-longitude NAM structure. The light red range shows the minimum to maximum range of the LENS members. Reanalyses (barely visible) are from left to right: ERA-Interim, MERRA2 and JRA55. (l) JFM averaged NAM timescales at 500hPa. Reanalyses are from left to right: ERA5, ERA-Interim, MERRA2 and JRA55 and black dashed line = ERA5.

Figure 14. (a)-(d) The DJF NAO for (a) ERA5, (b) BCAM6, (c) BCAM6−ERA5 and (d) LENS−ERA5. (e) the *NMSE* of the spatial structure of the DJF NAO from 20N-90N: gray = CMIP5 and 6; red = LENS with the light red range that spans the figure indicating the minimum and maximum for the 40 members; green = BCAM6; blue = BWACCM6; green hatched = FCAM6; blue hatched = FWACCM6; black from left to right = ERA5, ERA-Interim, MERRA2, JRA55 and the range of 36 years segments of ERA20C and 20CR. (f) The percentage of variance explained by the NAO with the same ordering as (e). (g)-(l) are as (a)-(f) but for the JJA NAO.

Figure 15. Blocking frequency (% of days) for the NH between 30°N and 75°N. (a)-(d) show DJF ERA5 climatology, BCAM6 climatology, BCAM6−ERA5 and LENS−ERA5, respectively. Gray shading in (c) and (d) indicate where ERA5 lies within the model ensemble member distribution. (e)-(h) are as (a)-(d) but for JJA. (i) and (j) *NMSE* (see legend) and SVR (circles) for DJF and JJA, respectively. The light red range that spans (i) and (j) depicts the minimum to maximum range for LENS and the reanalyses (barely visible) are ordered from left to right: ERA-Interim, MERRA2 and JRA55. When the unconditional bias corresponds to a bias in the spatial mean of more than 10% of the ERA5 spatial mean, the red (blue) SVR circles depict positive (negative) biases.
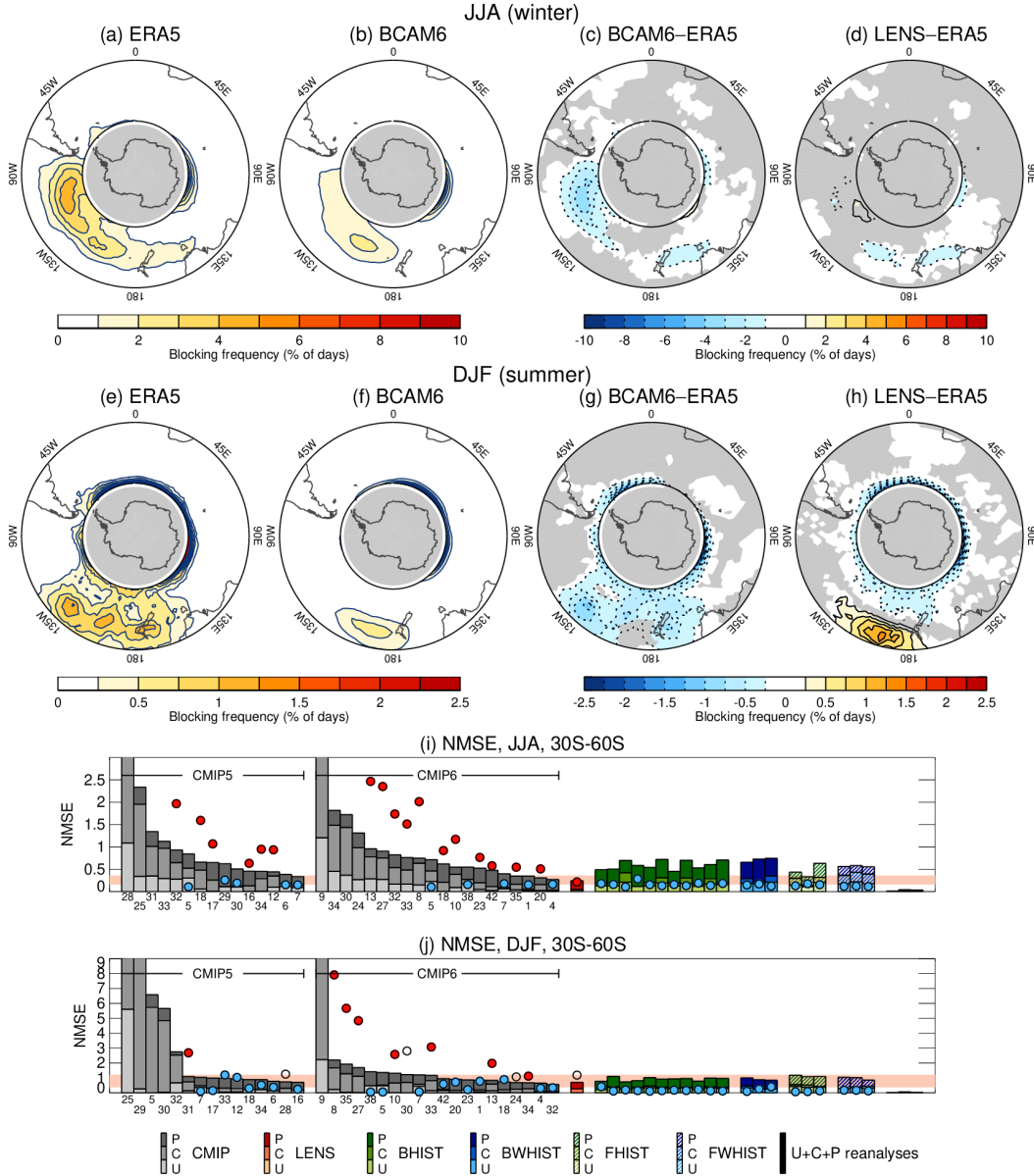
Figure 16. Blocking frequency (% of days) for the SH between 30°S and 60°S. (a)-(d) show the JJA ERA5 climatology, BCAM6 climatology, BCAM6−ERA5 and LENS−ERA5, respectively. Gray shading in (c) and (d) indicates where ERA5 lies within the model ensemble member distribution. (e)-(h) are as (a)-(d) but for DJF. (i) and (j) *NMSE* (see legend) and SVR (circles) for JJA and DJF, respectively. The light red range that spans (i) and (j) indicates the minimum to maximum ranges for LENS and the reanalyses are ordered from left to right: ERA-Interim, MERRA2 and JRA55. When the unconditional bias corresponds to a bias in the spatial mean of more than 10% of the ERA5 spatial mean, the red (blue) SVR circles depict positive (negative) biases.
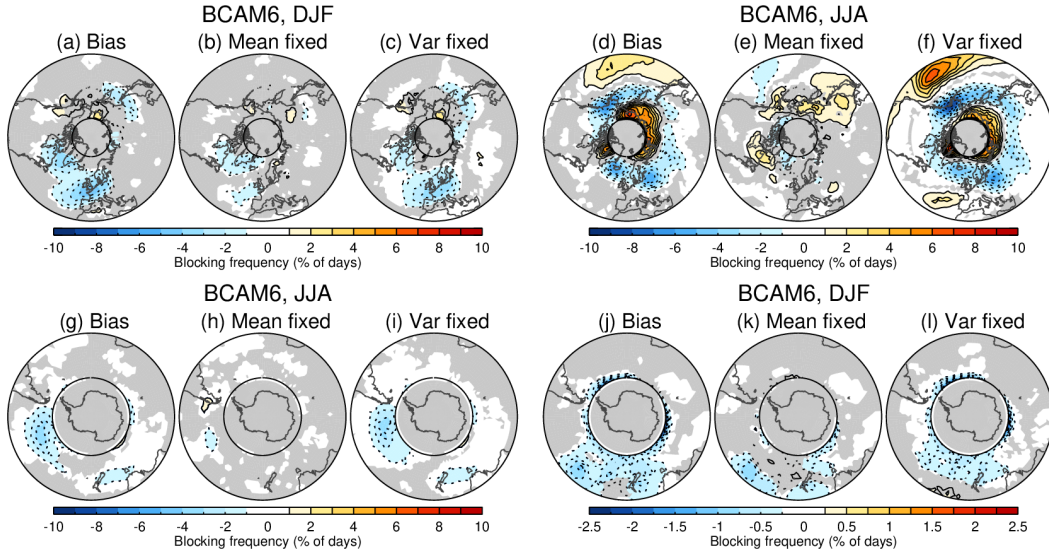
Figure 17. (a)-(c) shows a decomposition of the NH DJF blocking bias in BCAM6 (a) into a contribution that is present when the seasonal mean climatology of zg is replaced by that of ERA5 (b) and a contribution that is present when the deviations from the seasonal mean climatology of zg are replaced by those of ERA5 (c) i.e., (b) shows the bias that would be present if the mean state were improved but the variability left unchanged and (c) shows the bias that would be present if the variability were improved but the mean state left unchanged. (d)-(f), (g)-(i) and (j)-(k) are as (a)-(c) but for the NH during JJA, the SH during JJA and the SH during DJF, respectively.
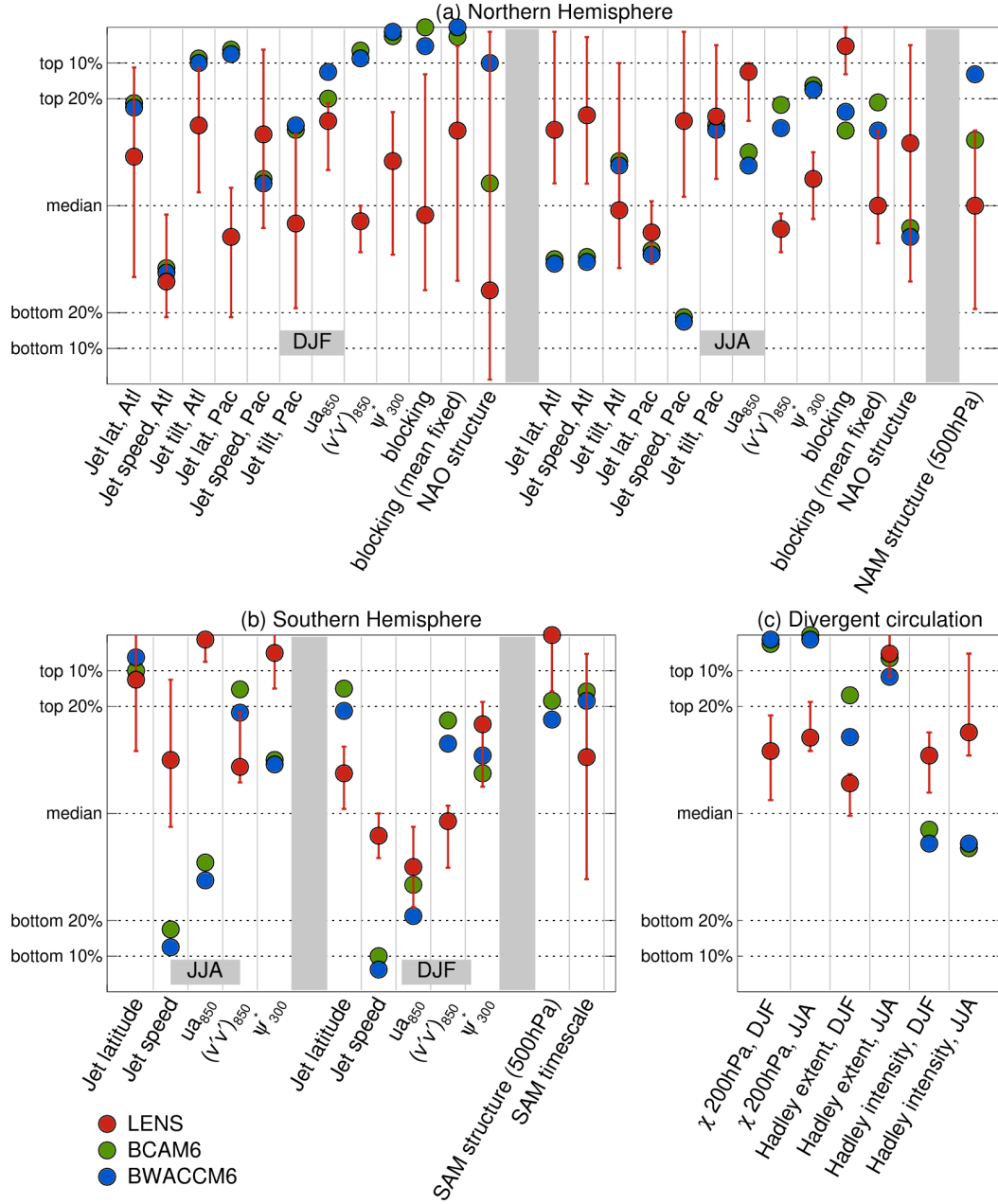
Figure 18. A summary of the representation of various (a) NH features, (b) SH features and (c) aspects of the global divergent circulation in CESM2-CAM6 (BCAM6), CESM2-WACCM6 (BWACCM6), CESM1 (LENS) as compared to the distribution of CMIP5 and CMIP6 models combined. Red point and range = the LENS ensemble mean and the range from the worst to best individual ensemble member. Green = BCAM6 ensemble mean. Blue = BWACCM6 ensemble mean. Y-axis displays the ranking of these CESM configurations among the CMIP5 and CMIP6 models combined, expressed as a percentile from the worst at the bottom to the best at the top. The CMIP distribution consists of 77 models for monthly fields, 43 models for daily *ua* and *va* metrics and 35 models for daily *zg* metrics.
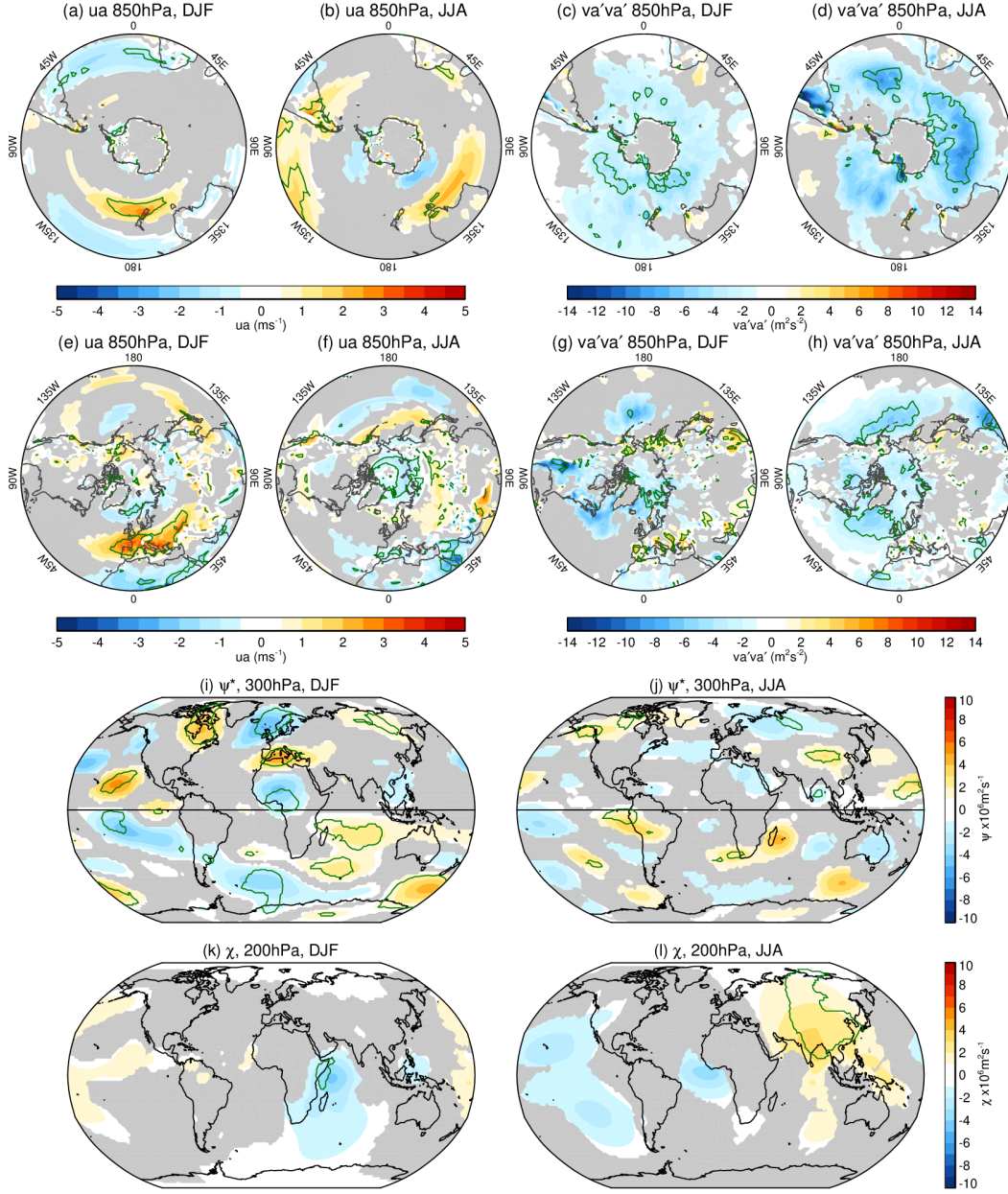
Figure A1. CMIP6 ensemble mean bias relative to ERA5. Gray shading depicts regions where less than 75% of the models agree on the sign of the bias and green contour shows where more than 95% of the models agree on the sign of the bias. (a), (b), (e) and (f) show 850hPa *ua*. (c) (d), (g) and (h) show 850hPa 10-day high pass eddy meridional wind variance. (i) and (j) show 300hPa eddy stream function and (k) and (l) show 200hPa velocity potential.
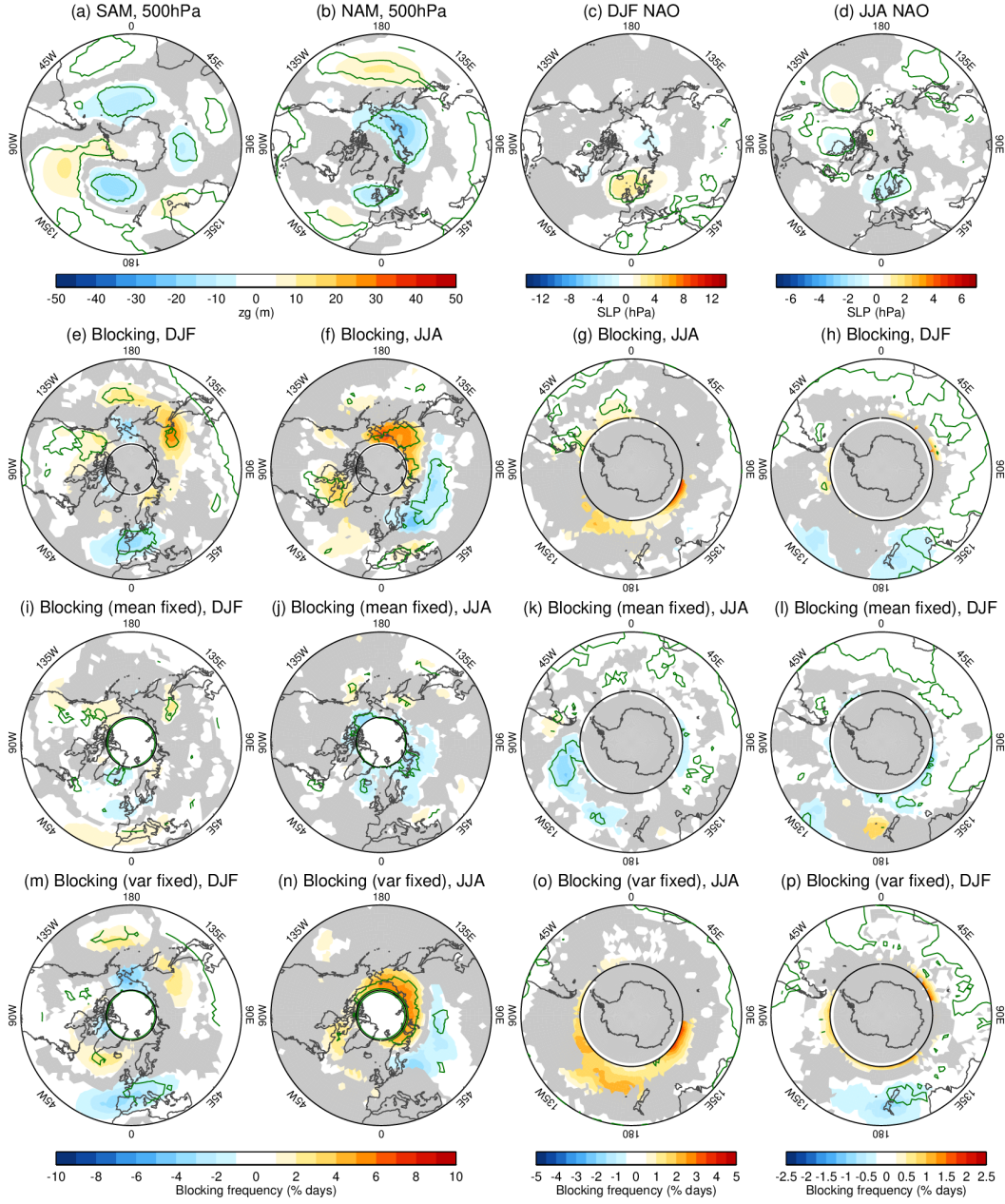
Figure A2. CMIP6 ensemble mean biases relative to ERA5 for variability metrics. Gray shading depicts regions where less than 75% of the models agree on the sign of the bias and green contour shows where more than 95% of models agree on the sign of the bias. (a) and (b) show the 500hPa zg structure associated with the SAM and NAM. (c) and (d) show the winter and summer NAO. (e)-(h) show the representation of blocking. (i)-(l) are as (e)-(h) but replacing each models climatology with that of ERA5 before calculating the blocking statistics. (m)-(p) are as (e)-(h) but replacing each models transient variability (deviations from climatology) with that of ERA5 before calculating the blocking statistics.

Appendix A Common biases in the CMIP6 models

For many of the features discussed, systematic biases in CMIP5 have been assessed in the various studies cited in the text. Given that such studies are not yet available for the newer CMIP6 archive, we provide an overview of systematic biases in the CMIP6 models in the features considered here in Figs. A1 and A2. In these figures, the ensemble mean bias relative to ERA5 is shown by the color shading where more than 75% of models agree on the sign of the bias. Green contours outline regions where more than 95% of models agree on the sign of the bias. CESM2-CAM6 and CESM2-WACCM6 have not been included in this analysis.

For the SH westerlies, there is a clear systematic bias toward the westerlies being too strong to the South of New Zealand and the easterly trade winds to be too strong over much of the hemisphere in DJF (Fig. A1a). In JJA, there is strong model agreement on westerlies that are too strong in the low latitude Pacific and more than 75% consensus on the westerlies being too strong to the South of Australia. For the NH westerlies, the major systematic bias across the CMIP6 models is the westerlies that are too strong over Europe with easterly biases to the south over North Africa in DJF (Fig. A1e). These accompany an anticyclonic bias centered over the Mediterranean (Fig. A1i).

For the storm tracks, as measured by the 850hPa 10-day high pass meridional wind variance, there is a systematic tendency toward storm tracks that are too weak in both hemispheres and all seasons (Figs. A1 and S14 c,d,g and h). This is particularly true downstream of the Andes and the Rockies during winter (Figs. A1d and g). This lack of meridional wind variance was also present in CESM1 but has now been improved in CESM2.

For the upper level divergent flow, there is less consensus among the models on the bias relative to ERA5 but there is strong agreement on too much divergence over the Indian Ocean in DJF (Fig. A1k), as well as convergence over the tropical Atlantic and divergence over Asia that are too weak during JJA (Fig. A1l).

For the variability, there is a strong consensus on the three localized positive centers of action in the zg SAM structure in the South Pacific, Southern Indian Ocean and South Atlantic to be too weak (Fig. A2a) and for the negative anomaly in zg of the NAM structure too be too strong over Northern Russia, as found in CESM2 (Fig. A2b). Models agree on a positive slp bias in the DJF NAO structure in the North Sea and a negative bias in the same region in the summer (Figs. A2c and d).

For blocking, models systematically underestimate European blocking during DJF (Fig. A2e) and underestimate JJA blocking over Eastern Europe/Western Russia, while producing too much blocking over Eastern Russia (Fig. A2f). However, these blocking biases are predominantly associated with a mean state bias that limits the ability of synoptic anticyclones to reverse the zg gradient, as opposed to there being a systematic lack of persistent anticyclones in these regions (see the decomposition in Figs. A2i, j, m and n). In the major blocking centers of the SH, there is no model agreement on systematic biases in JJA (Fig. A2g), while there is reasonable model agreement on a lack of blocking in the New Zealand sector and western Pacific during DJF (Fig. A2h).

Acknowledgments

References

Adam, O., Grise, K. M., Staten, P., Simpson, I. R., Davis, S. M., Davis, N. A., ... Ming, A. (2018). The TropD software package (v1): standardized methods for calculating tropical-width diagnostics. *Geosci. Model Dev.*, *11*, 4339–4357.

Ambaum, M. H. P., Hoskins, B. J., & Stephenson, D. B. (2001). Arctic Oscillation or North Atlantic Oscillation? *J. Clim.*, *14*, 3495–3507.

Anstey, J. A., Davini, P., Gray, L. J., Woollings, T. J., Butchard, N., Cagnazzo, C., ... Yang, S. (2013). Multi-model analysis of Norhtern Hemisphere winter blocking: Model biases and the role of resolution. *J. Geophys. Res.*, *118*, 3956–3971.

Baldwin, M. P., Stephenson, D. B., Thompson, D. W. J., Dunkerton, T. J., Charlton, A. J., & O'Neill, A. (2003). Stratospheric Memory and Skill of Extended-Range Weather Forecasts. *Science*, *301*, 636–640.

Barnes, E. A., Dunn-Sigouin, E., Masato, G., & Woollings, T. (2013). Exploring recent trends in Northern Hemisphere blocking. *Geophys. Res. Lett.*, *41*, 638–644.

Barnes, E. A., & Polvani, L. (2013). Response of the Midlatitude Jets, and of Their Variability, to Increased Greenhouse Gases in the CMIP5 Models. *J. Clim.*, *26*, 7117–7135.

Beljaars, A. C. M., Brown, A. R., & Wood, N. (2004). A new parametrization of turbulent orographic form drag. *Q. J. R. Meteorol. Soc.*, *130*, 1327–1347.

Benedict, J. J., Clement, A. C., & Medeiros, B. (2019). Atmospheric Blocking and Other Large-Scale Precursor Patterns of Landfalling Atmospheric Rivers in the North Pacific: A CESM2 Study. *J. Geophys. Res. Atm.*, *124*, 11330-11353.

Berrisford, P., Hoskins, B. J., & Tyrlis, E. (2007). Blocking and Rossby Wave Breaking on the Dynamical Tropopause in the Southern Hemisphere. *J. Atmos. Sci.*, *64*, 2881–2898.

Biastoch, A., Böning, C. W., Schwarzkopf, F. U., & Lutjeharms, J. R. E. (2009). Increase in Agulhas leakage due to poelward shift of Southern Hemisphere westerlies. *Nature*, *462*, 495–499.

Bladè, I., Fortuny, D., van Oldernborgh, G. J., & Liebmann, B. (2012). The summer North Atlantic Oscillation in CMIP3 models and related uncertainties in projected summer drying in Europe. *J. Geophys. Res.*, *117*, D166104.

Bogenschutz, P. A., Gettelman, A., Hannay, C., Larson, V. E., Neale, R. B., Craig, C., & Chen, C.-C. (2018). The path to CAM6: coupled simulations with CAM5.4 and CAM5.5. *Geosci. Model Dev.*, *11*, 235-255.

Bogenschutz, P. A., Gettelman, A., Morrison, H., Larson, V. E., Craig, C., & Schanen, D. P. (2013). Higher-Order Turbulence Closure and Its Impact on Climate Simulations in the Community Atmosphere Model. *J. Clim.*, *26*, 9655–9676.

Boville, B. A., & Gent, P. R. (1998). The NCAR Climate System Model, version 1. *J. Clim.*, *11*, 1115–1130.

Boyle, J. S. (2006). Upper level atmospheric stationary waves in the twentieth century climate of the Intergovernmental Panel on Climate Change simulations. *J. Geophys. Res.*, *111*, D14101.

Bracegirdle, T. J., Shuckburgh, E., Sallee, J.-B., Wang, Z., Meijers, A. J. S., Burneau, N., ... Wilcox, L. J. (2013). Assesssment of surface winds over the Atlantic, Indian and Pacific Ocean sectors of the Southern Ocean in CMIP5 models: historical bias, forcing response, and state dependence. *J. Geophys. Res.*, *118*, 1-16.

Broccoli, A. J., & Manabe, S. (1992). The Effects of Orography on Midlatitude Northern Hemisphere Dry Climates. *J. Clim.*, *5*, 1181–1201.

Brunner, L., Schaller, N., Anstey, J., Sillmann, J., & Steiner, A. K. (2018). Dependence of present and future European temperature extremes on the location of atmospheric blocking. *Geophys. Res. Lett.*, *45*, 6311-6320.

Buehler, T., Raible, C. C., & Stocker, T. F. (2011). The relationship of winter season North Atlantic blocking frequencies to extreme cold or dry spells in the ERA-40. *Tellus Ser. A.*, *63*, 212-222.

Byrne, N. J., & Shepherd, T. G. (2018). Seasonal Persistence of Circulation Anomalies in the Southern Hemisphere Stratosphere and Its Implications for the Troposphere. *J. Clim.*, *31*, 3467–3483.

Byrne, N. J., Shepherd, T. G., Woollings, T., & Plumb, R. A. (2016). Annular modes and apparent eddy feedbacks in the Southern Hemisphere. *Geophys. Res. Lett.*, *43*, 3897–3902.

C3S. (2019). *Copernicus Climate Change Service (C3S): ERA5: Fifth generation of ECMWF atmospheric reanalyses for the global climate.* Copernicus Climate Change Service Climate Data Store.

Caballero, R. (2008). Hadley cell bias in climate models linked to extratropical eddy stress. *Geophys. Res. Lett.*, *35*, L18709.

Capotondi, A., Deser, C., Phillips, A. S., Okumura, Y., & Larson, S. M. (2020). ENSO and Pacific Decadal Variability in the Community Earth System Model Version 2. *J. Adv. Mod. Earth Sys.*, *In Press.*

Ceppi, P., Hwang, Y.-T., Frierson, D. M. W., & Hartmann, D. L. (2012). Southern Hemisphere jet latitude biases in CMIP5 models linked to shortwave cloud forcings. *Geophys. Res. Lett.*, *39*, L19708.

Chang, E. K. M., Guo, Y., & Xia, X. (2012). CMIP5 multimodel ensemble projection of storm track change under global warming. *J. Geophys. Res.*, *117*, D23118.

Chang, E. K. M., Guo, Y., Xia, X., & Zheng, M. (2013). Storm-Track Activity in IPCC AR4/CMIP3 Model Simulations. *J. Clim.*, *26*, 246–260.

Chang, E. K. M., Lee, S., & Swanson, K. L. (2002). Storm Track Dynamics. *J. Clim.*, *15*, 2163–2183.

Cohen, J., Frei, A., & Rosen, R. D. (2005). The Role of Boundary Conditions in AMIP-2 Simulations of the NAO. *J. Clim.*, *18*, 973–981.

Collins, W. D., Rasch, P. J., Boville, B. A., Hack, J. J., McCaa, J. R., Williamson, D. L., & Briegleb, B. P. (2006). The Formulation and Atmospheric Simulation of the Community Atmosphere Model Version 3 (CAM3). *J. Clim.*, *19*, 2144–2161.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., . . . Worley, S. J. (2011). The twentieth century reanalysis project. *Quart. J. Roy. Meteor. Soc.*, *137*, 1–28.

Curtis, P. E., Ceppi, P., & Zappa, G. (2020). Role of the Mean State for the Southern Hemispheric Jet Stream Response to CO2 Forcing in CMIP6 models. *Env. Res. Lett.*, *In Press*. doi: https://doi.org/10.1088/1748-9326/ab8331

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., . . . Strand, W. G. (2019). The Community Earth System Model 2 (CESM2). *J. Adv. Mod. Earth Sys.*, *12*. doi: 10.1029/2019MS001916

D'Andrea, F., Tibaldi, S., Blackburn, M., Boer, G., Deque, M., Dix, M. R., . . . Williamson, D. (1998). Northern Hemisphere atmospheric blocking as simulated by 15 atmospheric general circulation models in the period 1979-1988. *Clim. Dyn.*, *14*, 385–407.

Davini, P., & D'Andrea, F. (2016). Northern Hemisphere Atmospheric Blocking Representation in Global Climate Models: Twenty Years of Improvements? *J. Clim.*, *29*, 8823–8840.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., . . . Vitart, F. (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *QJRMS*, *137*, 553–597. doi: 10.1002/qj.828

Delcambre, S. C., Lorenz, D. J., Vimont, D. J., & Martin, J. E. (2013). Diagnosing Northern Hemisphere Jet Portrayal in 17 CMIP3 Global Climate Models: Twentieth-Century Intermodel Veriability. *J. Clim.*, *26*, 4910–4929.

Doblas-Reyes, F., Deque, M., Valero, F., & Stephenson, D. B.   (1998).   North Atlantic wintertime intraseasonal variability and its sensitivity to GCM horizontal resolution. *Tellus*, *50A*, 573-595.

Dunn-Sigouin, E., & Son, S.-W.   (2013).   Northern Hemisphere blocking frequency and duration in the CMIP5 models. *J. Geophys. Res.: Atmospheres*, *118*, 1179-1188.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Mod. Dev.*, *9*, 1937–1958.

Fyfe, J. C., & Saenko, O. A.   (2006).   Simulated changes in the extratropical Southern Hemisphere winds and currents. *Geophys. Res. Lett.*, *33*, L06701.

Gelaro, R., McCarty, W., Surez, M. J., Todling, R., Molod, A., Takacs, L., . . . Zhao, B. (2017). The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *J. Clim.*, *30*, 5419–5454.

Gent, P. R. (2016). Effects of southern hemisphere wind changes on the meridional overturning circulation in models. *Geophys. Res. Lett.*, *33*, L06701.

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., . . . Zhang, M. (2011). The Community Climate System Model Version 4. *J. Clim.*, *23*, 4973–4991.

Gerber, E. P., Baldwin, M. P., Akiyoshi, H., Austin, J., Bekki, S., Braesicke, P., . . . Smale, D. (2010). Stratosphere-troposphere coupling and annular mode variability in chemistry-climate models. *J. Geophys. Res.*, *115*, D00MN06.

Gerber, E. P., Polvani, L. M., & Ancukiewicz, D. (2008). Annular mode time scales in the intergovernmental panel on climate change fourth assessment report models. *Geophys. Res. Lett.*, *35*, L22707. doi: 10.1029/2008GL035712

Gerber, E. P., & Vallis, G. K. (2007). Eddy-Zonal Flow Interactions and the Persistence of the Zonal Index. *J. Atmos. Sci.*, *64*, 3296–3311.

Gerber, E. P., Voronin, S., & Polvani, L. M. (2008). Testing the Annular Mode Autocorrelation Time Scale in Simple Atmospheric General Circulation Models. *Mon. Weath. Rev.*, *136*, 1523–1536.

Gettelman, A., Mills, M. J., Kinnison, D. E., Garcia, R. R., Smith, A. K., Marsh, D. R., . . . Randel, W. J.   (2019).   The Whole Atmosphere Community Climate Model Version 6 (WACCM6). *J. Geophys. Res.: Atm*, *124*. doi: 10.1029/2019JD030943

Gettelman, A., & Morrison, H.   (2015).   Advanced Two-Moment Bulk Microphysics for Global Models. Part 1: Off-Line Tests and Comparison with Other Schemes. *J. Clim.*, *28*, 1268–1287.

Gillett, N. P., & Fyfe, J. C.   (2013).   Annular mode changes in the CMIP5 simulations. *Geophys. Res. Lett.*, *40*. doi: doi:10.1002/grl.50249

Gillett, N. P., Kell, T. D., & Jones, P. D.   (2006).   Regional climate impacts of the Southern Annular Mode. *Geophys. Res. Lett.*, *33*, L23704.

Golaz, J.-C., Larson, V. E., & Cotton, W. R.   (2002).   A PDF-Based Model for Boundary Layer CLouds. Part 1: Method and Model Description. *J. Atmos. Sci.*, *59*, 3540–3551.

Gong, D., & Wang, S. (1999). Definition of the Antarctic Oscillation index. *Geophys. Res. Lett.*, *26*, 459-462.

Gong, H., Wang, L., Chen, W., Chen, X., & Nath, D. (2017). Biases of the wintertime Arctic Oscillation in CMIP5 models. *Env. Res. Lett.*, *12*, 014001.

Greeves, C. Z., Pope, V. D., Stratton, R. A., & Martin, G. M.   (2007).   Representation of Northern Hemisphere winter storm tracks in climate models. *Clim. Dyn.*, *28*, 683–702.

Hack, J. H., Boville, B. A., Briegleb, B. P., Kiehl, J. T., Rasch, P. J., & Williamson, D. L.   (1993).   *Description of the NCAR Community Climate Model (CCM2)* (Tech. Rep.).   NCAR.   Retrieved from https://opensky.ucar.edu/

islandora/object/technotes:149

Hannachi, A., Barnes, E. A., & Woollings, T. (2013). Behaviour of the winter North Atlantic eddy-driven jet stream in the CMIP3 integrations. *Clim. Dyn.*, *41*, 995-1007.

Hannachi, A., Woollings, T., & Fraedrich, K. (2012). The North Atlantic jet stream: a look at preferred positions, paths and transitions. *Q. J. Roy. Met. Soc.*, *138*, 862-877.

Held, I. M., Ting, M., & Wang, H. (2002). Northern Winter Stationary Waves: Theory and Modeling. *J. Clim.*, *15*, 2125–2144.

Hertwig, E., von Storch, J.-S., Handorf, D., Dethloff, K., Fast, I., & Krismer, T. (2015). Effect of horizontal resolution on ECHAM6-AMIP performance. *Clim. Dyn.*, *45*, 185-211.

Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., ... Zhang, Q. (2018). Historical (1750–2014) anthropogenic emissions of reactive gases and aerosols from the community emissions data system (ceds). *Geoscientific Model Development*, *11*(1), 369–408. doi: 10.5194/gmd-11-369-2018

Hoskins, B. J., & Hodges, K. I. (2005). A New Perspective on Southern Hemisphere Storm Tracks. *J. Clim.*, *18*, 4108–4129.

Hurrell, J. W. (1995). Decadal Trends in the North Atlantic Oscillation: Regional Temperatures and Precipitation. *Science*, *269*, 676–679.

Hurrell, J. W., Hack, J. J., Shea, D., Caron, J. M., & Rosinski, J. (2008). A New Sea Surface Temperature and Sea Ice Boundary Dataset for the Community Atmosphere Model. *J. Clim.*, *21*, 5145–5153.

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., ... Marshall, S. (2013). The Community Earth System Model: A Framework for Collaborative Research. *Bull. Amer. Meteor. Soc.*, *94*, 1339-1360.

Iqbal, W., Leung, W.-N., & Hannachi, A. (2018). Analysis of the variability of the North Atlantic eddy-driven jet stream in CMIP5. *Clim. Dyn.*, *51*, 235-247.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., ... Vertenstein, M. (2014). The Community Earth System Model (CESM) Large Ensemble Project. *Bull. Amer. Meteor. Soc.*, *96*, 1333–1349.

Kidston, J., & Gerber, E. P. (2010). Intermodel variability of the poleward shift of the austral jet stream in the CMIP3 integrations linked to biases in 20th century climatology. *Geophys. Res. Lett.*, *37*, L09708. doi: 10.1029/2010GL042873

Kiehl, J. T., & Gent, P. R. (2004). The Community Climate System Model, version Two. *J. Clim.*, *17*, 3666-3682.

Kiehl, J. T., Hack, J. J., Bonan, G. B., Boville, B. B., Williamson, D. L., & Rasch, P. J. (1998). The National Center for Atmospheric Research Community Climate Model: CCM3. *J. Clim.*, *11*, 1131–1149.

Kim, W. M., Yeager, S., Chang, P., & Danabasoglu, G. (2018). Low frequency North Atlantic climate variability in the Community Earth System Model Large Ensemble. *J. Clim.*, *31*, 787-813.

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., ... Takahashi, K. (2015). The jra-55 reanalysis: General specifications and basic characteristics. *J. Met. Soc. Japan.*, *93*, 5–48.

Kravtsov, S. (2017). Pronounced differences between observed and CMIP5-simulated multidecadal climate variability in the twentieth century. *Geophys. Res. Lett.*, *44*, 5749-5757.

Kwon, Y.-O., Camacho, A., Martinez, C., & Seo, H. (2018). North Atlantic winter eddy-driven jet and atmospheric blocking variability in the Community Earth System Model version 1 Large Ensemble simulations. *Clim. Dyn.*, *51*, 3275-3289.

Lamarque, J.-F., Bond, T. C., Eyring, V., Granier, C., Heil, A., Klimont, Z., ...

van Vuuren, D. P. (2010). Historical (18502000) gridded anthropogenic and biomass burning emissions of reactive gases and aerosols: methodology and application. *Atmospheric Chemistry and Physics*, *10*(15), 7017–7039.

Lamarque, J. F., Kyle, G. P., Meinshausen, M., Riahi, K., Smith, S. J., van Vuuren, D. P., . . . Vitt, F. (2011). Global and regional evolution of short-lived radiatively active gases and aerosols in the Representative Concentration Pathways. *Climatic Change*, *109*, 191-212.

Lefebvre, W., Goosse, H., Timmermann, R., & Fichefet, T. (2004). Influence of the Southern Annular Mode on the sea ice-ocean system. *J. Geophys. Res.*, *109*, C09005.

Leith, C. E. (1975). Climate response and fluctuation dissipation. *J. Atmos. Sci.*, *32*, 2022-2026.

L'Heureux, M. L., & Thompson, D. W. J. (2006). Observed Relationships between the El Niño-Southern Oscillation and the Extratropical Zonal-Mean Circulation. *J. Clim.*, *19*, 276–287.

Lorenz, D. J., & Hartmann, D. L. (2001). Eddy-Zonal Flow Feedback in the Southern Hemisphere. *J. Atmos. Sci.*, *58*, 3312-3327.

Marshall, J., & Speer, K. (2012). Closure of the meridional overturning circulation through Southern Ocean upwelling. *Nat. Geosc.*, *5*, 171–180.

Masato, G., Hoskins, B. J., & Woollings, T. (2013a). Wave-Breaking Characteristics of Northern Hemisphere Winter Blocking: A Two-Dimensional Approach. *J. Clim.*, *26*, 4535–4549.

Masato, G., Hoskins, B. J., & Woollings, T. (2013b). Winter and Summer Northern Hemisphere Blocking in CMIP5 Models. *J. Clim.*, *26*, 7044–7059.

McHugh, M. J., & Rogers, J. C. (2005). Multi-model representation of the North Atlantic Oscillation in the 20th and 21st centuries. *Geophys. Res. Lett.*, *32*, L21713.

Meehl, G. A., Shields, C., Arblaster, J. M., Annamalai, H., & Neale, R. (2020). Seasonal mean monsoon simulations in CESM2. *J. Adv. Mod. Earth Sys.*, *submitted*.

Meehl, G. A., Yang, D., Arblaster, J. M., Bates, S. C., Rosenbloom, N., Neale, R., . . . Danabasoglu, G. (2019). Effect of model resolution, physics, and coupling on Southern Hemisphere storm tracks in CESM1.3. *Geophys. Res. Lett.*, *46*, 12408-12416.

Meinshausen, M., Smith, S. J., Calvin, K., Daniel, J. S., Kainuma, M. L. T., Lamarque, J.-F., . . . van Vuuren, D. P. P. (2011). The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. *Climatic Change*, *109*, 213-241.

Miller, R. L., Schmidt, G. A., & Shindell, D. T. (2006). Forced annular variations in the 20th century Intergovernmental Panel on Climate Change Fourth Assessment Report models. *J. Geophys. Res.*, *111*, D18101.

Murphy, A. H. (1988). Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient. *Mon. Weath. Rev.*, *116*, 2417–2424.

Neale, R. B., Richter, J., Park, S., Lauritzen, P. H., Vavrus, S. J., Rasch, P. J., & Zhang, M. (2013). The Mean Climate of the Community Atmosphere Model (CAM4) in Forced SST and Fully Coupled Experiments. *J. Clim.*, *26*, 5150-5168.

Patterson, M., Bracegirdle, T., & Woollings, T. (2019). Southern Hemisphere Atmospheric Blocking in CMIP5 and Future Changes in the Australia-New Zealand Sector. *Geophys. Res. Lett.*, *46*. doi: https://doi.org/10.1029/2019GL083264

Pfahl, S., & Wernli, H. (2012). Quantifying the relevance of atmospheric blocking for co-located temperature extremes in the Norhtern Hemisphere on (sub-)daily time scales. *Geophys. Res. Lett.*, *39*, L12807.

Pithan, F., Shepherd, T. G., Zappa, G., & Sandu, I. (2016). Climate model biases in jet streams, blocking and storm tracks resulting from missing orographic drag.

Geophys. Res. Lett., *43*, 7231-7240.

Poli, P., Hersbach, H., Dee, D. P., Berrisford, P., Simmons, A. J., Vitart, F., . . . Fisher, M. (2016). Era-20c: An atmospheric reanalysis of the twentieth century. *J. Clim.*, *29*, 4083–4097.

Richter, J. H., Sassi, F., & Garcia, R. R. (2010). Toward a Physically Based Gravity Wave Source Parameterization in a General Circulation Model. *J. Atmos. Sci.*, *67*, 136–156.

Rodwell, M. J., & Hoskins, B. J. (2000). Subtropical Anticyclones and Summer Monsoons. *J. Clim.*, *14*, 3192–3211.

Russell, J. L., Stouffer, R. J., & Dixon, K. W. (2006). Intercomparison of the Southern Ocean Circulations in IPCC Coupled Model Control Simulations. *J. Clim.*, *19*, 4560–4575.

Sardeshmukh, P. B., & Hoskins, B. J. (1988). The Generation of Global Rotational Flow by Steady Idealized Tropical Divergence. *J. Atmos. Sci.*, *45*, 1228–1251.

Scaife, A. A., Woollings, T., Knight, J., Martin, G., & Hinton, T. (2010). Atmospheric Blocking and Mean Biases in Climate Models. *J. Clim.*, *23*, 6143–6152.

Schaller, N., Sillmann, J., Anstey, J., Fischer, E. M., Grams, C. M., & Russo, S. (2018). Influence of blocking on Northern European and Western Russian heatwaves in large climate model ensembles. *Env. Res. Lett.*, *13*, 054015.

Seager, R., Liu, H., Henderson, N., Simpson, I., Kelley, C., Shaw, T., . . . Ting, M. (2014). Causes of Increasing Aridification of the Mediterranean Region in Reponse to Rising Greenhouse Gases. *J. Clim.*, *27*, 4655–4676.

Sen Gupta, A., & England, M. H. (2006). Coupled Ocean-Atmosphere-Ice Response to Variations in the Southern Annular Mode. *J. Clim.*, *19*, 4457–4486.

Shaw, T. A., Baldwin, M., Barnes, E. A., Caballero, R., Garfinkel, C. I., Hwang, Y.-T., . . . Voigt, A. (2016). Storm track processes and the opposing influences of climate change. *Nat. Geosc.*, *9*, 656–665.

Shaw, T. A., Perlwitz, J., & Weiner, O. (2014). Troposphere-stratosphere coupling: Links to North Atlantic weather and climate, including their representation in CMIP5 models. *J. Geophs. Res. Atm.*, *119*, 5864–5880.

Shaw, T. A., & Voigt, A. (2015). Tug of war on summertime circulation between radiative forcing and sea surface warming. *Nat. Geosc.*, *8*, 560–567.

Sillmann, J., Croci-Maspoli, M., Kallache, M., & Katz, R. W. (2011). Extreme cold winter temperatures in Europe under the Influence of North Atlantic atmospheric blocking. *J. Clim.*, *24*, 5899-5013.

Silverman, B. W. (1981). Using Kernel Density Estimates to investigate multimodality. *J. R. Statist. Soc. Series B*, *43*, 97-99.

Simpson, I. R., Deser, C., McKinnon, K. A., & Barnes, E. a. (2018). Modeled and Observed Multidecadal Variability in the North Atlantic Jet Stream and Its Connection to Sea Surface Tempereatures. *J. Clim.*, *31*, 8313–8338.

Simpson, I. R., Hitchcock, P., Shepherd, T. G., & Scinocca, J. F. (2011). Stratospheric variability and tropospheric annular-mode timescales. *Geophys. Res. Lett.*, *38*, L20806.

Simpson, I. R., & Polvani, L. M. (2016). Revisiting the relationship between jet position, forced response, and annular mode variability in the southern midlatitudes. *Geophys. Res. Lett.*, *43*, 2896–2903.

Simpson, I. R., Seager, R., Ting, M., & Shaw, T. A. (2016). Causes of change in Northern Hemisphere winter meridional winds and regional hydroclimate. *Nat. Clim. Change*, *6*, 65–70.

Simpson, I. R., Shepherd, T. G., Hitchcock, P., & Scinocca, J. F. (2013). Southern Annular Mode Dynamics in Observations and Models. Part 2: Eddy Feedbacks. *J. Clim.*, *26*, 5220–5241.

Simpson, I. R., Yeager, S. G., McKinnon, K. A., & Deser, C. (2019). Decadal predictability of late winter precipitation in western Europe through an ocean-jet

stream connection. *Nat. Geosc.*, *12*, 613–621.

Sinclair, M. R. (1996). A Climatology of Anticyclones and Blocking for the Southern Hemisphere. *Mon. Weath. Rev.*, *124*, 245–263.

Son, S., Gerber, E. P., Perlwitz, J., Polvani, L. M., Gillett, N., Seo, K.-H., ... Yamashita, Y. (2010). Impact of Stratospheric Ozone on the Southern Hemisphere Circulation Changes : A Multimodel Assessment. *J. Atmos. Sci.*, *115*, 1–55.

Stephenson, D. B., & Pavan, V. (2003). The North Atlantic Oscillation in coupled climate models: a CMIP1 evalution. *Clim. Dyn.*, *20*, 381-399.

Stephenson, D. B., Pavan, V., Collins, M., Junge, M. M., & Quadrelli, R. (2006). North Atlantic Oscillation response to transient greenhouse gas forcing and the impact on European winter climate: a CMIP2 multi-model assessment. *Clim. Dyn.*, *27*, 401-420.

Stoner, A. M. K., Hayhoe, K., & Wuebbles, D. J. (2009). Assessing General Circulation Model Simulations of Atmospheric Teleconnection Patterns. *J. Clim.*, *22*, 4348–4372.

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, *92*, 485-498.

Thompson, D. W. J., & Solomon, S. (2002). Interpretation of Recent SOuthern Hemisphere Climate Change. *Science*, *296*, 895–899.

Thompson, D. W. J., & Wallace, J. M. (1998). The Arctic Oscillation signature in the wintertime geopotential height and temperature fields. *Geophys. Res. Lett.*, *25*, 1297-1300.

Thompson, D. W. J., & Wallace, J. M. (2000). Annular modes in the extratropical circulation. Part 1: Month-to-month variability. *J. Clim.*, *13*, 1000-1016.

Tibaldi, S., & Molteni, F. (1990). On the operational predictability of blocking. *Tellus*, *42A*, 343-365.

Tim, N., Zorita, E., Emels, K.-C., Schwarzkoft, F. U., Biastoch, A., & Hünicke, B. (2019). Influence of position and strength of westerlies and trades on Agulhas leakage and South Benguela Upwelling. *Earth Syst. Dynam. Discuss.*, *In Review*. doi: https://doi.org/10.5194/esd-2019-16

Trenberth, K. E., & Mo, K. C. (1985). Blocking in the Southern Hemisphere. *Mon. Weath. Rev.*, *113*, 3-21.

Trigo, R. M., Trigo, I. F., DaCamara, C. C., & Osborn, T. J. (2004). Climate impact of the European winter blocking episodes from the NCEP/NCAR Reanalyses. *Clim. Dyn.*, *23*, 17-28.

Ulbrich, U., Pinto, J. G., Kupfer, J., Leckebusch, G. C., Spangehl, T., & Reyers, M. (2008). Changing Northern Hemisphere Storm Tracks in an Ensemble of IPCC Climate Change Simulations. *J. Clim.*, *21*, 1669–1679.

Ummenhofer, C. C., McIntosh, P. C., Pook, M. J., & Risbey, J. S. (2013). Impact of Surface Forcing on Southern Hemisphere Atmospheric Blocking in the Australia-New Zealand Sector. *J. Clim.*, *26*, 8476–8494.

van Marle, M. J. E., Kloster, S., Magi, B. I., Marlon, J. R., Daniau, A.-L., Field, R. D., ... van der Werf, G. R. (2017). Historic global biomass burning emissions for CMIP6 (BB4CMIP) based on merging satellite observations with proxies and fire models (1750–2015). *Geoscientific Model Development*, *10*(9), 3329–3357. doi: 10.5194/gmd-10-3329-2017

van Niekerk, A., Scinocca, J. F., & Shepherd, T. G. (2017). The Modulation of Stationary Waves, and Their Response to Climate Change, by Parameterized Orographic Drag. *J. Atmos. Sci.*, *74*, 2557–2574.

Vial, J., & Osborn, T. J. (2012). Assessment of atmosphere-ocean general circulation model simulations of winter northern hemisphere atmospheric blocking. *Clim. Dyn.*, *39*, 95-112.

Wang, H., & Ting, M. (1999). Seasonal Cycle of the Climatological Stationary Waves in the NCEP-NCAR Reanalysis. *J. Atmos. Sci.*, *56*, 3892–3919.

Wang, X., Li, J., Sun, C., & Liu, T. (2017). NAO and its relationship with the Northern Hemisphere mean surface temperature in CMIP5 simulations. *J. Geophys. Res. Atmos.*, *122*, 4202-4227.

Washington, W. M. (1982). *Documentation of the Community Climate Model (CCM), Version 0* (Tech. Rep.). NCAR.

White, R. H., Hilgenbrink, C., & Sheshadri, A. (2019). The Importance of Greenland in setting the Northern Preferred Position of the North Atlantic Eddy-driven Jet. *Geophys. Res. Lett.*, *46*. doi: 10.1029/2019GL084780

Wilcox, L. J., Charlton-Perez, A. J., & Gray, L. J. (2012). Trends in Austral jet position in ensembles of high- and low-top CMIP5 models. *J. Geophys. Res.*, *117*, D13115.

Williamson, D. L. (1983). *Description of the NCAR Community Climate Model (CCM0B)* (Tech. Rep.). NCAR. Retrieved from `https://opensky.ucar.edu/islandora/object/technotes:317`

Williamson, D. L. (1995). *Proc. AMIP Scientific Conf.* Monterey, CA: World Climate Research Program.

Williamson, D. L., Kiehl, J. T., Ramanathan, V., Dickinson, R. E., & Hack, J. J. (1987). *Description of NCAR Community Climate Model (CCM1)* (Tech. Rep.). NCAR. Retrieved from `https://opensky.ucar.edu/islandora/object/technotes:400`

Wills, R. C. J., White, R. H., & Levine, X. J. (2019). Northern Hemisphere Stationary Waves in a Changing Climate. *Current Climate Change Reports*. doi: https://doi.org/10.1007/s40641-019-00147-6

Woollings, T. (2010). Dynamical influecnes on European climate: an uncertain future. *Phil. Trans. Roy. Soc. Lon.*, *368*, 3733-3756.

Woollings, T., Barriopedro, D., Methven, J., Son, S.-W., Martius, O., Harvey, B., ... Seneviratne, S. (2018). Blocking and its Response to Climate Change. *Current Climate Change Reports*, *4*, 287-300.

Woollings, T., & Blackburn, M. (2011). The North Atlantic Jet Stream under Climate Change and Its Relation to the NAO and EA Patterns. *J. Clim.*, *25*, 886–902.

Woollings, T., Hannachi, A., & Hoskins, B. (2010). Variability of the North Atlantic eddy-driven jet stream. *Q. J. Roy. Met. Soc.*, *136*, 856-868.

Zappa, G., Shaffrey, L. C., & Hodges, K. I. (2013). The Ability of CMIP5 Models to SImulate North Atlantic Extratropical Cyclones. *J. Clim.*, *26*, 5379–5396.

Zuo, J.-Q., Li, W.-J., & Ren, H.-L. (2013). Representation of the Arctic Oscillation in the CMIP5 models. *Adv. Clim. Change Res.*, *4*, 242–249.