The Climate-system Historical Forecast Project: Do stratosphere-resolving models make better seasonal climate predictions in boreal winter?

Amy H. Butler[1,2], Alberto Arribas[3], Maria Athanassiadou[3], Johanna Baehr[4], Natalia Calvo[5], Andrew Charlton-Perez[6], Michel Déqué[7], Daniela I.V. Domeisen[8], Kristina Fröhlich[9], Harry Hendon[10], Yukiko Imada[11], Masayoshi Ishii[11], Maddalen Iza[5], Alexey Yu. Karpechko[12], Arun Kumar[13], Craig MacLachlan[3], William J. Merryfield[14], Wolfgang A. Müller[15], Alan O'Neill[6], Adam A. Scaife[3], John Scinocca[14], Michael Sigmond[14], Timothy N. Stockdale[16], and Tamaki Yasuda[17]

[1]University of Colorado/Cooperative Institute for Research in Environmental Sciences, Boulder, CO, USA
[2]National Oceanic and Atmospheric Administration/Earth Systems Research Laboratory/Chemical Sciences Division, Boulder, CO, USA
[3]Met Office Hadley Centre, Exeter, UK
[4]Institute of Oceanography, Center for Earth System Research and Sustainability (CEN), Universität Hamburg, Germany
[5]Universidad Complutense de Madrid, Madrid, Spain
[6]University of Reading, Reading, UK
[7]Météo-France/Centre National de Recherches Meteorologiques/CNRS-GAME, Toulouse, France
[8]GEOMAR Helmholtz Centre for Ocean Research Kiel/University of Kiel, Kiel, Germany
[9]Deutscher Wetterdienst (DWD), Offenbach, Germany
[10]Bureau of Meteorology, Melbourne, Australia
[11]Meteorological Research Institute, Japan Meteorological Agency, Ibaraki, Japan
[12]Finnish Meteorological Institute, Helsinki, Finland
[13]National Oceanic and Atmospheric Administration/National Weather Service/Climate Prediction Center, College Park, MD, USA
[14]Canadian Centre for Climate Modelling and Analysis/Environment Canada, Victoria, BC, Canada
[15]Max Planck Institute for Meteorology, Hamburg, Germany
[16]European Centre for Medium-Range Weather Forecasts, Reading, UK
[17]Japan Meteorological Agency, Tokyo, Japan

Corresponding author email address: amy.butler@noaa.gov

Abstract

Using an international, multi-model suite of historical forecasts from the World Climate Research Programme (WCRP) Climate-system Historical Forecast Project (CHFP), we compare the seasonal prediction skill in boreal wintertime between models that resolve the stratosphere and its dynamics ("high-top") and models that do not ("low-top"). We evaluate hindcasts that are initialized in November, and examine the model biases in the stratosphere and how they relate to boreal wintertime (Dec-Mar) seasonal forecast skill. We are unable to detect more skill in the high-top ensemble-mean than the low-top ensemble-mean in forecasting the wintertime North Atlantic Oscillation, but model performance varies widely. Increasing the ensemble size clearly increases the skill for a given model. We then examine two major processes involving stratosphere-troposphere interactions (the El Niño-Southern Oscillation/ENSO and the Quasi-biennial Oscillation/QBO) and how they relate to predictive skill on intra-seasonal to seasonal timescales, particularly over the North Atlantic and Eurasia regions. High-top models tend to have a more realistic stratospheric response to El Niño and the QBO compared to low-top models. Enhanced conditional wintertime skill over high-latitudes and the North Atlantic region during winters with El Niño conditions suggests a possible role for a stratospheric pathway.

Keywords: seasonal prediction, stratosphere-troposphere coupling, stratosphere, El Niño-Southern Oscillation, Quasi-biennial Oscillation, North Atlantic Oscillation

1. Introduction

The predictive skill of extratropical climate on intraseasonal to interannual time scales has historically been very low. Interactions between the ocean, ice, atmosphere, and land on intermediate timescales can be complex, and many models do not capture these relationships (National Research Council 2010; Kim et al. 2012; Smith et al. 2012). In an effort to assess current seasonal prediction capabilities, the World Climate Research Programme (WCRP) proposed the Climate-system Historical Forecast Project (CHFP), in which historical forecasts, or hindcasts, from international operational forecast centers and other research centers could be compared and evaluated (Kirtman and Pirani 2009).

One component of the atmosphere that models historically have had difficulty simulating is the stratosphere. Because stratospheric variability can have significant impacts on surface climate on timescales of weeks to months during seasons when the stratosphere is dynamically coupled to the troposphere (Gerber et al. 2012; Kidston et al. 2015), incorporating a well-resolved stratosphere into forecasting models is one of the most promising ways to enhance intra-seasonal to seasonal prediction skill (Baldwin et al. 2003; National Research Council 2010; Smith et al. 2012). Models with a well-resolved stratosphere, i.e. those with a higher model top and vertical resolution in the stratosphere, seem to better simulate stratospheric variability and stratosphere-troposphere coupling (Osprey et al. 2013; Charlton-Perez et al. 2013) and have better stratospheric and tropospheric skill compared to models with a lower model top on timescales of 3-4 weeks (Roff et al. 2011). Initializing models within a few days of a major stratospheric event can enhance predictive skill at the surface during the weeks following the event (Kuroda 2008; Gerber et al. 2009; Marshall and Scaife 2010;

Sigmond et al. 2013; Tripathi et al. 2014, 2015), and forcing the stratosphere towards observed values improves the simulation of the wintertime tropospheric circulation (Douville 2009; Scaife et al. 2005).

However, it is still not clear if initializing and resolving the stratosphere results in higher seasonal forecast skill in the Northern Hemisphere wintertime. In a seasonal forecast, models are initialized at the beginning of the season and predictions are made for the subsequent months. Models cannot predict sub-monthly stratospheric variability (e.g., major Sudden Stratospheric Warmings/SSWs) at these lead times because the information used to initialize the model generally does not provide useful knowledge about the tropospheric wave processes that drive these events beyond ~7-20 days. Thus, any predictive information gained from having a well-resolved stratosphere is likely due to the ability of the model to capture (a) those stratospheric processes most related to slow-varying atmospheric processes and teleconnections, and (b) the downward transfer of those stratospheric signals to tropospheric climate, which tends to occur on intra-seasonal to seasonal timescales.

Two elements that influence the stratosphere and tend to persist throughout a season are: tropical sea surface temperature anomalies, in the form of the El Niño-Southern Oscillation (ENSO), and the Quasi-biennial Oscillation (QBO), a quasi-periodic oscillation of tropical stratospheric winds. Both ENSO and the QBO have seasonal to interannual timescales and are associated with changes in extratropical surface climate, particularly over the North Atlantic and Eurasian regions. The stratosphere plays an important role in communicating these tropically-forced teleconnections to the extratropical and polar latitudes. Incorporating realistic dynamics of the stratosphere may improve the simulation of these processes and their impacts on extratropical climate on seasonal timescales (Cagnazzo and Manzini 2009; Hardiman et al. 2012; Boer and Hamilton 2008).

Here we seek to compare and quantify differences in the seasonal prediction skill between models with a resolved stratosphere (so-called “high-top” models) and without a resolved stratosphere (“low-top” models). While other studies have examined the stratospheric circulation and its role in seasonal predictive skill in low-top models (Maycock et al. 2011), as well as its role in decadal to long-term predictions (Scaife et al. 2012, 2014a), this study is the first intercomparison of historical seasonal forecasts made by a large international suite of both high- and low-top state-of-the-art forecasting models (**Table I**; Section 2). We focus on model hindcasts that are initialized on November 1st and run through the Northern Hemisphere wintertime (December-March) when the stratosphere is strongly coupled to the troposphere and has known influences on surface climate. We evaluate the model biases in the climatology and variability of the stratosphere, and relate them to biases in the tropospheric circulation and skill (section 3). We compare the predictive capability of the models for the North Atlantic Oscillation (NAO), the dominant mode of extratropical variability associated with shifts in the tropospheric mid-latitude North Atlantic jet (section 4). Lastly we consider whether high-top or low-top models better represent stratospheric teleconnections from the tropics to the extratropics, associated with the ENSO (section 5) and the QBO (section 6).

2. CHFP Dataset and Methodology

Table I provides a list of the coupled atmosphere-ocean models participating in the CHFP, along with the number of ensemble members, model resolution and model top, years of the model hindcast run, seasonal range of the forecast from November-March, data output levels above the surface, the dataset used to initialize the atmosphere, and the simulation of the ocean and sea ice. Many modeling centers provide nearly 30 years of historical forecast data, and multiple ensemble members, allowing for large sample sizes. The atmosphere, ocean, and sea ice in the models are initialized on or around November 1st (the CHFP dataset also includes hindcasts initialized around May 1st but here we focus on the boreal winter season). While some of the models create ensemble members by initializing runs from late October into early November, here we mainly consider predictive skill in December through March, so any additional information from forecasts initialized after Nov 1st should not influence the results. The data set is hosted at the Centro de Investigaciones del Mar y la Atmosfera (CIMA) and is publicly available for download at: http://chfps.cima.fcen.uba.ar/

In this study we use monthly-mean data (about half of the models also provide daily data in the stratosphere). Here we do not explicitly examine the role of extreme stratospheric variability that occurs on timescales of a week or less (i.e., major SSWs or strong polar vortices) and its relationship to slowly varying boundary conditions such as ENSO. Seasonal forecast models cannot generally gain skill from simulating individual SSWs (since models can only forecast SSWs at lead times of 10-20 days), but the increased probability of SSWs during certain winters is associated with improved predictive skill at the surface (e.g., Domeisen et al. 2015; Scaife et al. 2015).

To be considered a "high-top" or stratosphere-resolving model (bolded and italicized in **Table I**), the model must have a vertical domain extending to 1 hPa (~50 km) or higher, and have at least 15 model levels between the tropopause and 1 hPa. We note that while a higher model top and better stratospheric vertical resolution presumably allow a better representation of wave processes and thus better representation of stratospheric dynamics (Shaw and Perlwitz 2010; Charlton-Perez et al. 2013), it's not clear that this condition alone is sufficient for better simulation of stratospheric processes (Shaw et al. 2014). For example, even most high-top models are unable to simulate an internally-generated QBO. Still, using model lid height to classify the models into two groups does generally separate those models with weak polar stratospheric variability from those with stronger, more realistic variability (see **Figure 1d**). Note that while several models offer data output at a wide range of pressure levels, other models do not provide any data above 200 hPa; in these cases we cannot evaluate the model stratosphere and its relation to skill beyond knowing the model lid height. These models are included in analysis of surface variables (i.e., **Figures 4** and **6**) but not in the analysis of stratospheric variables.

While a few modeling centers provide a high-top and low-top version of the same model (e.g., ARPEGE, CMAM, and GloSea4), overall this suite of models is an "ensemble of opportunity". In other words, the high-top models and the low-top models may not only be different in terms of the model lid height, but also in terms of model physics, resolution, and other parameterizations. These differences can make it difficult to attribute differences in skill directly to the inclusion of a more realistic stratosphere.

We compare high- and low- top versions of the same model, in relation to the ensemble-mean differences, where possible.

To calculate skill and evaluate model relationships against the observations, we use the ECMWF ERA-interim reanalysis dataset (Dee et al. 2011). Comparing to other reanalysis datasets produced very similar results (not shown). We evaluate monthly-mean and seasonal-mean Northern Hemisphere wintertime (November-March) fields from November 1979 to March 2012.

Individual model, high-top, and low-top ensemble-mean skill are evaluated following Becker et al. (2014). For model forecasts *F(s, j, n, k)*, where *s* is the spatial (gridpoint) index, *j* is the time index (years), *n* is the model ensemble member *n*=1…N for N total ensemble members per model, and *k* is the model *k*=1…K for K total models, the ensemble mean (EM) for model *k* is given by:

$$F_{ens}(s,j,k) = \sum_n F(s,j,n,k)/N \qquad (1)$$

Likewise, the high-top (F_H) and low-top (F_L) ensemble-means are formed by averaging those model ensemble-means that qualify for each category, e.g.:

$$F_H(s,j) = \sum_k F_{ens}(s,j,k)/K \qquad (2)$$

where *$F_{ens}(s,j,k)$* is the ensemble-mean of the *k*-th high-top model out of K total high-top models. Note that individual model EMs (F_{ens}) are equally weighted within the high-top and low-top EMs (i.e., models with more members are not given more weight). In addition, note that each F_{ens} has a different time period (**Table I**), so to create the high-top and low-top EMs, each F_{ens} has been padded with missing data values if necessary to ensure every record runs from 1979-2012 prior to averaging into F_H or F_L. This means that the high-top and low-top EM time series are based on few models at the beginning and ends of the record. Using a somewhat shorter time period to eliminate periods with few models has little effect on our results.

To create anomalies, the ensemble-mean model climatology (for the individual model period) is then removed:

$$F'_{ens}(s,j,k) = F_{ens}(s,j,k) - \{F_{ens}(s,k)\} \tag{3a}$$

and

$$F'_H(s,j) = F_H(s,j) - \{F_H(s)\} \tag{3b}$$

where {} is the mean, or climatology, over the forecast period 1979-2012. The removal of the EM climatology allows an *a posteriori* removal of systematic errors, giving bias-corrected anomalies (Peng et al. 2002). Seasonal-means are created by averaging anomalies over 3 month periods.

The anomaly correlation coefficient (ACC) is used as a measure of skill, and is defined as:

$$ACC = \frac{\sum_s \sum_j \frac{w_s F'_H(s,j) O'(s,j)}{W}}{\left\{\sum_s \sum_j \frac{w_s F'_H(s,j)^2}{W} * \sum_s \sum_j \frac{w_s O'^{(s,j)^2}}{W}\right\}^{\frac{1}{2}}} \tag{4}$$

where *O'(s, j)* is the ERA-interim anomaly for every grid space *s* and year *j* (for the same years as the forecast), w_s is a weight that accounts for the area of each gridpoint, and *W* is the sum of w_s over all gridpoints and timesteps. The ACC values fall between -1 and +1, where +1 is a perfect forecast and 0 is the average score of a random forecast.

For calculations of the NAO pattern and time series, the following procedure is adapted from *Doblas-Reyes et al* (2003). The NAO is calculated as the first empirical orthogonal function (EOF) of 20-90°N and 90°W-60°E sea level pressure anomalies (SLP). As a way to consider the mid-tropospheric circulation we also calculate the Northern Annular Mode (NAM) as the first EOF of zonal-mean 20-90°N geopotential height anomalies at 500 hPa. For ERA-interim reanalysis, the EOF patterns are calculated based on detrended, deseasonalized DJF monthly anomalies weighted by the square root of cosine of latitude. The index time series is then found by projecting the EOF onto the unweighted gridded monthly anomalies, averaging over December-February (DJF), and standardizing the resulting time series. For models, the EOF pattern is calculated based on concatenated DJF monthly timeseries of all ensemble members for a given model. The NAO/NAM index time series is found by projecting the anomalies for each individual model ensemble member onto the model's EOF pattern (each model ensemble-mean climatology is used to create anomalies for each model). Projecting the anomalies onto the reanalysis-based EOF pattern instead did not significantly impact the results. All individual model NAO/NAM time series are then averaged together to make the model ensemble-mean time series, which is then standardized. These time series are correlated to the ERA-interim NAO/NAM index time series to get the skill of the ensemble-mean forecast. For the high-top and low-top ensemble-means, individual model-means of the NAO time series are averaged together, so that each model-mean is weighted equally.

We also approximate a "station-based" NAO index by subtracting mean sea level pressure (SLP) anomalies between the Azores Islands (38°N, 26°W) and Reykjavik, Iceland (64°N, 22°W). The SLP time series are normalized before finding the difference. The DJF seasonal-mean of the NAO index for each ensemble member is found, and then the ensemble members are averaged together to get a model-mean NAO value. Low-top and high-top ensemble-means are created by averaging each model-mean NAO value, so each model is weighted equally.

We define ENSO phases using the ERSST.V3B 'Oceanic Niño index' (ONI) calculated over the Niño-3.4 region (5°S-5°N, 170-120°W) from the National Center for Environmental Prediction (NCEP) Climate Prediction Center (CPC). El Niño and La Niña winters are classified following the NCEP/CPC convention: events must exceed the +0.5°C or -0.5°C threshold, for El Niño and La Niña respectively, for a minimum of five consecutive overlapping seasons (NDJ, DJF, JFM, etc.). For the period 1979-2012, there are 10 El Niño years, 12 La Niña years, and 11 ENSO-neutral years (**Table II**); note though that different forecast systems cover varying lengths of this 33 year period (**Table I**). Because the historical forecast runs are initialized in November with observed sea surface temperatures, the forecast for each winter will closely correspond to the observed ENSO phase.

Likewise, given the approximate 28-month periodicity of the Quasi-biennial Oscillation, the tropical winds initialized in the model in November should persist through each forecast winter (Marshall and Scaife 2009). The QBO phases are defined based on the November 5°S-5°N zonal-mean zonal wind at the ERA-Interim reanalysis level of 44 hPa. The westerly QBO (WQBO) phase occurs when the zonal-mean zonal winds are greater than 5 m s^{-1}, and the easterly QBO (EQBO) phase occurs when the zonal-mean zonal winds are less than -5 m s^{-1}. There are 14 WQBO winters and 12 EQBO winters (**Table II**).

3. Model biases in the Stratosphere

We first consider the ability of each model to simulate the mean state of the stratospheric polar vortex during Northern Hemisphere (NH) winter. **Figure 1a** shows the 1979-2012 climatological 50-70°N zonal winds at 50 hPa from November to March for ERA-interim (black dots, with confidence interval about the mean given by a 2-tailed *t*-test), low-top models (dashed lines), and high-top models (solid lines). In general there is wide spread behavior in the models compared to the reanalysis, which shows the vortex strengthening from November to January, peaking around 20 m s^{-1}, and weakening from January to March. Three out of five of the low-top models[1] (MIROC5, CanCM3, and ARPEGE_z00l) and one of the seven high-top models[1] (ARPEGE_z00k) have a stratospheric vortex that is too weak for most of the winter. Both the low-top and high-top versions of CMAM have a polar vortex that is too strong by February. A number of forecast systems capture the strength and/or the evolution of the vortex accurately for most of the winter (MPI-ESM-LR, MPI-ESM-MR, CFS, ECMWF-s4, CanCM4, ECMWF-s4, and GloSea5). Overall, there is not a strong difference in the mean state of the stratospheric polar vortex between high- and low-top models, though more low-top models are biased, in agreement with Maycock et al (2011).

Differences between high-top and low-top models are more apparent in the variability of the stratosphere. **Figure 1b** shows the standard deviation of the polar vortex winds for individual models (calculated for individual members first and then averaged) compared to ERA-interim, and **Figure 1d** shows the same figure but for the high-top and low-top EMs. While the high-top models tend to closely simulate the

[1] With data available at 50 hPa

observed stratospheric variability and its evolution (with peak variance in February), the low-top models tend to have much lower stratospheric variability from January to March. The lack of stratospheric variability in low-top models has also been noted in coupled climate models (Charlton-Perez et al. 2013). There are some exceptions, however: the low-top models CanCM4 and CMAMlo have reasonable representations of polar vortex variability, whereas the two high-top models CFS and ARPEGE_z00k lack variability in February and March.

Due to the chaotic nature of extreme polar stratospheric variability in wintertime, models generally cannot predict extreme stratospheric variations more than 10-20 days ahead of time (though during certain winters, such as El Niño, there may be an increased probability of these events; see Section 5). The information initialized in the models (in this case, around early November of each year) provides some positive predictive skill for the first few weeks of the model run, after which predictability decreases through the rest of the season (**Figure 1c**) as the forecast relaxes to the model climatology. This skill is slightly higher for the high-top model EM, but not significantly different from the low-top EM skill. Positive, non-zero skill is maintained through January, but only after January is the model skill higher than the persistence forecast.

How might model biases in the stratospheric jet relate to predictive skill at the surface? Biases in the strength of the stratospheric zonal circulation may be associated with biases in the tropospheric mid-latitude jet location (Gerber and Polvani 2009), which seems to hold true in the CHFP models particularly for the East Pacific tropospheric jet location[2] (**Figure 2a**, correlation significant at $p<0.001$) and to a weaker extent for the Atlantic tropospheric jet location (**Figure 2b**, correlation significant at $p<0.10$). Biases in the strength of the stratospheric zonal circulation are also associated with biases in the *strength* of the Atlantic tropospheric jet (35-55°N, 300-360°E) (**Figure 2d**, correlation significant at $p<0.001$), but not the East Pacific tropospheric jet (**Figure 2c**). However, little relationship is found between the bias in the polar vortex and the forecasting skill of the tropospheric jet during DJF or the skill of the NAO index (not shown). Still, biases in the strength and variability of the polar vortex may be related to biases in the persistence of tropospheric jet variability (Gerber and Polvani 2009)- with a stronger and more variable polar vortex associated with more persistent tropospheric jets- and thus to the time scales that stratospheric variability couples to the troposphere (see also **Figure 7**).

We briefly consider the ability of the models to simulate the mean state and variability of the tropical stratosphere (**Figure 3**). Because few models resolve the gravity wave spectrum in the tropics, which also drives the QBO (Sato and Dunkerton 1997; Ern et al. 2014), the standard deviation of the 10°S-10°N zonal winds at 50 hPa tends to be much too low in models compared to ERA-interim after initialization in November (**Figure 3b**). GloSea5, ECMWF-s4, and MPI-ESM-MR internally simulate the QBO, but only GloSea5 (Scaife et al. 2014b) and MPI-ESM-MR have tropical wind variances that agree well with observations past December. In general, high-top models have slightly higher variance in tropical winds than low-top models. Despite the lack of tropical variability, the skill of the models in capturing the tropical winds is quite high

[2] Here, the tropospheric jet position is defined as the location of the maximum 850 hPa zonal wind speed between 15-75°N in the Atlantic (300-360°E) basin and E. Pacific (210-240°E) basin.

and persistent (**Figure 3a**). Most models maintain the state of the initialized tropical winds (which is dominated by the state of the QBO) due to slow radiative relaxation rates in the tropical lower stratosphere (Haynes 1998). Only ARPEGE_z00l does not maintain the mean state of the tropical winds past November. Section 6 will further consider whether the QBO-like tropical winds in the models influence extratropical climate on seasonal timescales.

4. Skill in forecasting the NAO

The North Atlantic Oscillation (NAO) is the regional manifestation of the hemispheric Northern Annular Mode (NAM; also known as the Arctic Oscillation/AO) and the dominant pattern of climate variability in the Northern Hemisphere. The NAO represents an oscillation in atmospheric mass between the Azores subtropical high and the Icelandic polar low and is associated with latitudinal shifts in the position of the North Atlantic storm track. These fluctuations are coupled to variability in the strength of the stratospheric polar vortex in boreal winter and have significant influences on climate in Eurasia, eastern North America, and Greenland. Thus there is considerable interest in improving the seasonal forecast skill of the NAO.

Given that most of the NAO's variability is due to internal atmospheric dynamics and feedbacks, prediction of the NAO at time scales longer than a couple of weeks is difficult. Nonetheless, some fraction of wintertime NAO variability may be externally forced on longer timescales (Keeley et al. 2009); for example, by variability in Arctic sea ice (e.g., García-Serrano et al 2015) or Eurasian snow cover extent (e.g., Cohen and Jones 2011). Ensemble members that better capture observed external forcings like snow cover extent have been shown to have higher NAO skill (e.g., Riddle et al 2013). It has been hypothesized that the stratosphere provides the pathway for an external forcing to persist through the winter season and impact the NAO and Eurasian surface climate (e.g., Ineson and Scaife 2009). Here we use the CHFP suite of models to examine whether a more resolved stratosphere improves the predictive skill of the DJF-mean NAO index for forecasts initialized in early November.

Table III shows the skill of the ensemble-mean forecast for the DJF-mean NAO index (calculated using both the EOF and station-based methods described in Section 2) and the 500 hPa NAM skill. The observed NAO pattern using the EOF method is well simulated by the CHFP models, with spatial correlations exceeding 0.91 for all models (**Figure S1**). Nonetheless the forecast skill for both the NAO and the NAM is generally low and not significant at the 95% level, for most high-top and low-top models (see also probabilistic skill scores in **Table S1**, and error bars in **Figure S2**). For the NAO EOF-based index, the high-top model ensemble-mean has significant skill (r=0.45, $p<0.01$), while the low-top ensemble mean does not (r=0.32, $p<0.07$), but both ensemble-means have significant skill when considering the Azores-Iceland station-based NAO index (neither have significant skill for the 500 hPa NAM). Between models where model lid height and vertical resolution are the only difference (i.e., CMAM, ARPEGE, and GloSea4), the NAO/NAM skill is not better in the high-top versions, suggesting that a higher model top does little to improve NAO/NAM skill for the 1979-2012 period.

GloSea5 has significant skill for all three NAO/NAM indices, with correlation values more than double most other models, and performs skillfully particularly for upper tercile NAO events (**Table S1**). However, GloSea5 also has one of the shortest hindcast periods of 14 years (1996-2010), and the length of the hindcast period is associated with large sampling uncertainty (Kumar 2009; Shi et al. 2015). To test how often a correlation value of r=0.61 (the skill of the GloSea5 NAO station-based index) occurs by chance, we calculate correlation values between the NAO station-based index for individual model-means and the reanalysis using random consecutive 14-year periods from all the models. From this distribution of 234 14-year runs, we find that a correlation above 0.6 occurs only 5 times (~2%), which implies that the GloSea5 skill is unlikely to occur randomly (we note that the other 4 times occur using 14-year periods from the MPI-ESM-MR model). The GloSea5 model has demonstrated similarly high NAO skill in longer 20 year hindcasts, perhaps due to in part to enhanced ocean resolution, initialization of Arctic sea ice, and increased ensemble members (Scaife et al. 2014a). It is interesting to note that Scaife et al. (2015) demonstrate that the wintertime NAO forecast skill in GloSea5 vanishes when considering only ensemble members with an inactive stratosphere (i.e., no sudden warmings in a given winter). Their results suggest a key role for stratosphere-troposphere coupling in the high NAO skill for this model.

We also test the dependence of the NAO skill on the particular hindcast period. Riddle et al (2013) noted higher forecast skill for the wintertime NAO in the Climate Forecast System version 2 for the 1997-2010 period compared to the 1983-1996 period, and other studies have found a similar dependence on the hindcast period (Kang et al. 2014; Müller et al. 2005). We consider the skill of the station-based NAO index over the same time period as GloSea5 (1996/97-2009/10) in the other models (**Table III**; note some models do not extend to 2009/10, in which case the correlation was found from 1996 to the end of the record). Some models show weaker or even negative correlations during this short time period compared to the full period, but other models more than double their skill. In particular, MPI-ESM-MR has skill of r=0.71 (p<0.004) for the DJF NAO during the 1996-2010 time period.

The number of ensemble members in the ensemble-mean plays a significant role in the skill of DJF NAO forecasts (**Figure 4**). For the majority of models, as the number of ensemble members increase, the DJF NAO skill score also increases (by enhancing the signal to noise ratio) as noted in previous studies (Kumar and Hoerling 2000; Kharin et al. 2001; Riddle et al. 2013; Chen et al. 2013; Scaife et al. 2014a; DelSole et al. 2014; Eade et al. 2014). GloSea5 has a large number of ensemble members (24), which contributes to its high skill score. Improved skill for the DJF NAO index has been found for both ECMWF-s4 (Stockdale et al. 2015) and version 2 of the NOAA CFS (Riddle et al 2013) when ensemble members are substantially increased. Nonetheless **Figure 4** indicates that for a given number of ensemble members, certain models perform better than others. For example, for 7 ensemble members, the DJF NAO skill ranges from ~0.04-0.37, with MPI-ESM-MR performing superior (note though that correlations here are for the length of each model record, which varies from 14 to 33 years (**Table III**), and the time period of the forecast may also be relevant as previously discussed).

Our results suggest that the DJF-mean NAO skill is not significantly different between the high-top and low-top models when initialized in early November. This is true even when comparing high-top/low-top pairs, like CMAM, ARPEGE, and GloSea4 (in many cases the low-top model actually performs substantially better than the high-top model, though it depends on which index is used). However, it is possible that during years when there is strong anomalous forcing of the stratosphere, such as during El

Niño/QBO winters or winters when SSWs occur, the seasonal forecast skill may be improved in models that resolve stratospheric processes and associated coupling between the stratosphere and troposphere (e.g., Orsolini et al 2009). We explore this possibility in the next sections.

5. Impact of El Niño's stratospheric response on seasonal forecasting skill

The planetary wave trains that emanate from the tropical Pacific Ocean region due to anomalous sea surface temperatures (SSTs) associated with ENSO can drive the amplification of vertical wave propagation into the stratosphere. During El Niño winters, for example, a strengthening and eastward shift of the Aleutian low can be in observed in the North Pacific. This signal is hypothesized to enhance the wave flux into the stratosphere through linear interference (Smith et al. 2010; Fletcher and Kushner 2011). The breaking of these waves in the stratosphere warms and weakens the stratospheric polar vortex in both observations and models (Garfinkel and Hartmann 2008; García-Herrera et al. 2006; Hurwitz et al. 2014; Manzini et al. 2006), though the wintertime evolution of the response can depend on the location or strength of the maximum warming of tropical Pacific SSTs (Toniazzo and Scaife 2006; Calvo et al. 2015, in review).

During La Niña winters, in general the opposite sign anomalies occur (i.e., the vortex cools and strengthens) in the seasonal-mean over the long-term record. However, in the last two decades a large number of SSWs, which weaken and warm the vortex, have occurred during both La Niña and El Niño winters (Butler and Polvani 2011), so that the observed stratospheric response to La Niña is less clear. In addition, the model-mean stratospheric response to La Niña during the 1979-2012 time period is opposite to that observed, a finding that is left for further investigation. To simplify the interpretation, we focus on the model response to El Niño only. We focus on the North Atlantic- European region, where stratospheric variability has been shown to make a significant difference on the surface climate response over Greenland, Europe, and Asia during El Niño winters (Ineson and Scaife 2009; Butler et al. 2014; Domeisen et al. 2015).

Figure 5 shows the composite wind anomalies (50-80°N) for El Niño winters for the high-top (left) and low-top models (center), as well as ERA-interim reanalysis (right). Here, individual model members are composited for El Niño winters and then averaged together for each model to get the model-mean response. The model composite includes those models where data above 200 hPa is available (**Table I**; 7 high-top models and 5 low-top models); note also that each model time period covers different El Niño events. The reanalysis (**Fig 5c**) indicates that the polar vortex initially strengthens in early winter in response to El Niño, but then weakens from January through March. The observed response here is weaker than in studies using a longer observational record, perhaps because of an increase of central Pacific-type El Niños during this time period (e.g., Lee and McPhaden 2010), which have a more ambiguous impact on the stratosphere (Garfinkel et al. 2013; Iza and Calvo 2015); or because of the confounding influence of the QBO, discussed more below.

Both the high-top and low-top models show a similar evolution of zonal wind anomalies, but the weakening of the vortex in late winter is stronger and more significant at 50 hPa in the high-top models than the low-top models. While the anomalously strong vortex in early winter would be initialized in all models, the model must be able to simulate the propagation of waves associated with ENSO into the stratosphere to drive the weakening vortex in late winter. As shown in **Figure 1b**, the polar stratospheric variability in the low-top models is generally too low, which is also associated with less weakening of the vortex in El Niño winters (Jan-Mar) in the low-top models. Nevertheless, the negative tropospheric zonal-mean zonal wind anomalies in March are significant in both sets of models.

The ENSO-associated anomalies in the polar vortex can descend into the troposphere and affect mid-latitude weather via shifts in the mid-latitude jets, particularly in the North Atlantic region. **Figure 6** shows the skill of the high-top and low-top EMs for JFM mean sea level pressure anomalies over the North Atlantic region, for all winters, El Niño winters, and ENSO-neutral winters. The spatially-averaged anomaly correlation coefficient/ACC (equation (4)) for the region shown (25-85°N, 90°W-90°E) is given in the upper right corner of each panel. The highest skill over the North American-North Atlantic-European region occurs with the high-top EM during El Niño winters, with low skill in both model EMs during ENSO-neutral winters. Interestingly, the high-top EM has nearly double the ACC of the low-top EM during all winters and El Niño winters (note that while the ACC values are modest, they are likely still significant given the large effective sample size created by aggregating over a large spatial domain and multiple winters, e.g. Becker et al. 2014). Since skill is high near subtropical North America in all cases, the main differences arise from skill over the polar cap, Eurasia, and central Europe. For example, the skill over Greenland and the Arctic in the low-top EM is close to zero or negative, while the skill in the high-top EM is positive over much of the polar cap. Unfortunately, skill over Europe is negative in both model sets, possibly related to poor simulation of Gulf Stream processes (Danabasoglu et al. 2010) and associated Atlantic blocking (Scaife et al. 2011).

Higher skill for high-top models during El Niño winters is not as evident for individual high-top/low-top model pairs (**Figure S3**). Both ARPEGE and CMAM show higher skill in JF mean sea level pressure anomalies over the North Atlantic region in the low-top model compared to the high-top model (this is true for DJF anomalies as well). GloSea4 shows better skill in the high-top model for JFM anomalies, but there are only 4 El Niño winters during the 14-year hindcast period for the low-top version so sampling is likely an issue. These results are difficult to interpret given the few ensemble members and short hindcast periods of individual models.

The high-top EM therefore shows improvement in conditional skill for mean sea level pressure anomalies over the low-top model EM, which may in part be due to better simulation of stratospheric dynamics during El Niño winters. This argument is reasonable given that little difference in skill is observed in ENSO-neutral winters, and the improvement in skill in the high-top EM occurs predominantly over the polar cap, Greenland, and Eurasia, where stratosphere-troposphere coupling exerts its greatest impacts. The results in **Figure 5** suggest that the stratosphere responds to El Niño forcing more strongly in high-top models compared to low-top models, and this may contribute to better surface climate prediction in high-top models during El Niño (**Figure 6**). However, we can’t rule out the potential role of other factors, such as improved simulation of tropospheric/oceanic/sea-ice dynamics related to ENSO in the high-top models.

Improved simulation of the stratospheric circulation may also not be fully reflected in tropospheric skill due to lack of stratosphere-troposphere coupling in the models, particularly in the low-top models. **Figure 7** compares the December and January stratosphere-troposphere coupling in ERA-interim, the high-top EM, and the low-top EM. In December (left column) the zonal wind anomalies at 50 hPa (50-80°N) are highly correlated to stratospheric anomalies in November and to lower tropospheric anomalies from December to January (correlation values of r~0.4-0.6). Neither the high-top nor low-top models capture the persistence of the observed stratospheric anomalies from November to December. However, the observed correlation of December zonal wind anomalies at 50 hPa to January zonal wind anomalies at 850 hPa is near the median value of correlation for high-top members (less than a quartile of low-top members have a similar correlation, and the median correlation is much lower). In January the observed zonal wind anomalies in the stratosphere are also strongly coupled to the surface (r~0.7), but the high-top and low-top models show weaker coupling (less than a quartile of both low-top and high-top ensemble members have as high of correlation, though the median correlation value is higher in the high-top ensemble than the low-top member ensemble). The observed persistence of the coupling of January stratospheric anomalies into February near the surface is fairly well-simulated by both the high-top and low-top model ensembles. In summary, the coupling of the stratosphere to the troposphere in these models tends to be weaker than the one observed "realization"; though we note that the observed correlation does fall within the range of simulated correlations. However, the high-top models do tend to have higher correlation values (closer to the observed value) than the low-top models, suggesting that reducing the stratospheric biases, and consequently improving the stratosphere-troposphere coupling (Gerber and Polvani 2009), may increase the amount of skill at the surface gained from stratospheric variability.

The QBO may also play a role in the skill in **Figure 6**. For example, during this particular period from 1980-2012, 6 of the 10 El Niño winters occurred during the westerly QBO phase, while only 2 of the 10 El Niño winters occurred during the easterly QBO phase (**Table II**). Garfinkel and Hartmann (2010) find that the El Niño teleconnection is actually stronger during WQBO, which could be enhancing the skill. However, it is difficult to separate these external forcings given the small number of ENSO/QBO events during the hindcast period. In the next section, we consider the role of the QBO on seasonal forecasting skill, while keeping in mind the concurrence of these phenomena.

6. Impact of QBO on seasonal forecasting skill

The QBO is an oscillation of tropical stratospheric zonal winds from easterly to westerly with a periodicity of roughly 28 months. Like ENSO, it is a tropical phenomenon but can influence extratropical climate and the polar stratosphere by modulating wave propagation and breaking (Holton and Tan 1980; Baldwin et al. 2001). Because the QBO is both predictable and persistent, it is potentially useful for seasonal forecasting (Boer and Hamilton 2008; Scaife et al. 2014b; Marshall and Scaife 2009).

Figure 8 (right column) shows the observed difference in the zonal wind response between the westerly QBO (WQBO) phase and the easterly QBO (EQBO) phase for December through February. As expected, during the WQBO (EQBO), the Northern

Hemisphere stratospheric polar vortex tends to be stronger (weaker) than normal (e.g., Dunkerton and Baldwin 1991; Garfinkel and Hartmann 2007). These differences have the strongest amplitude and extension to the surface in January but persist the entire winter season.

As discussed in Section 3 and shown in **Figure 3**, few models participating in CHFP are able to simulate a QBO-like oscillation, but most models are able to persist the initialized tropical stratospheric winds for at least the first 2-3 months before the winds relax towards model climatology. **Figure 8** (left and middle columns) shows the composite QBO response (WQBO minus EQBO) for the high-top models and the low-top models. The high-top model response in December (one month after initialization) is quite similar to the observed response in the Northern Hemisphere, though with stronger extension to the extratropical surface than observed. The low-top models also have the same sign response as observed, but stronger than observed anomalous westerlies near 30°N and weaker than observed westerly anomalies in the lower stratosphere and troposphere near 60°N. By January (**Figure 8**, middle row), the NH polar stratospheric response in high-top models has weakened considerably, and in the low-top models has weakened and turned easterly at 50 hPa polewards of 60°N. By February (**Figure 8**, bottom row), the NH polar stratospheric response is also the opposite sign to the observed response in both the high-top and low-top models (though note the tropical zonal wind response persists through the entire winter in both model sets, as expected from **Figure 3**).

While high-top models seem to better represent the QBO effect in the NH wintertime polar stratosphere, particularly with lead times of a month or less, the extratropical sea level pressure response to QBO (WQBO minus EQBO) is weak and insignificant for both the high-top and low-top ensemble-means (not shown). This result can be inferred from the weak extension of the response to the extratropical surface in **Figure 8**, and is robust to the level used to define the QBO. Scaife et al (2014b) note that despite reasonable prediction of the QBO itself, many models still fail to capture the QBO teleconnection to the surface. The response may also be small because of concurrent mixed signals from ENSO (**Table II**). Clearly, better simulation of the QBO extratropical teleconnection to the surface is a potential avenue for future improvement in seasonal to interannual prediction.

7. Discussion and Conclusions

We have provided an overview of the CHFP models in the context of a high-top and low-top model comparison, and evaluated the biases and performance of these models in the NH wintertime stratosphere. The CHFP offers a unique opportunity to survey a large number of forecast models in regards to whether improved stratospheric representation improves seasonal prediction at the surface. We have shown that high-top models better simulate the observed wintertime polar stratospheric variance (**Figure 1d**). However, model biases in the stratosphere, which we find correlated to position (strength) of the East Pacific (Atlantic) tropospheric jet, vary widely by individual model and appear independent of model lid height (**Figure 2**). We also find that the DJF NAO skill depends more strongly on ensemble-member size (**Figure 4**) than on the model top (**Table III**).

So do stratosphere-resolving models make better seasonal predictions in boreal winter? For the CHFP suite of models, increases in skill for mean sea level pressure anomalies over the polar cap and Eurasian regions are seen for the high-top EM compared to the low-top EM for the period 1979-2012 (**Figure 6**). Most of this enhanced skill appears to come from winters with ENSO forcing. Without a forced change in the wave driving or propagation into the stratosphere, a model initialized in November will quickly move away from the observations towards model climatology, no matter how improved the stratosphere. High-top models are able to better simulate the stratospheric response to both ENSO and QBO (**Figures 5, 8**), which may improve surface skill even 2-3 months after initialization during El Niño years (**Figure 6**).

We caution though that these results are based on an ensemble of opportunity, with each model having different tropospheric and stratospheric representations regardless of model top. Comparing the three model pairs with only differences in model lid height and vertical resolution (CMAM, ARPEGE, and GloSea4) gives more ambiguous results (**Figure S3**; though the skill is based on much fewer ensemble members and fewer El Niño events in each case than the high-top EM). In addition, though the high-top models show stronger responses to ENSO and QBO in the stratosphere than the low-top models (**Figures 5, 8**), there is generally little difference in the response near the surface. The improved stratospheric representation in the high-top models corresponds to stronger coupling between the stratosphere and troposphere, but generally not as strong as observed (**Figure 7**).

The CHFP provides model hindcasts initialized on or around November 1st, which is early in the winter season and means that ENSO and its teleconnections must be well-simulated in the model to drive the appropriate tropospheric wave fluxes throughout the subsequent season. More skill might be gained by initializing the model in December or January during the peak of NH tropospheric wave driving, and considering the late winter response. In addition, Domeisen et al (2015) find the most improved skill over the North Atlantic-Eurasian region for El Niño winters with a SSW event (when only those model members that also forecast a SSW event are included), as SSWs appear to dominate the wintertime surface climate response in the North Atlantic-European region (Butler et al. 2014). The CHFP has daily zonal wind data at 10 hPa for 7 models, so connections between extreme stratospheric variations and higher skill during ENSO and non-ENSO winters could be examined in future work.

Most new forecasting models will increase or already have increased model lid height and vertical resolution. While this change will certainly improve stratospheric variance and representation of the ENSO and QBO stratospheric response, further skill in surface climate may be gained by increasing model members and improving stratospheric coupling to surface climate.

Acknowledgments

We acknowledge the WCRP/CLIVAR Working Group on Seasonal to Interannual Prediction (WGSIP) for establishing the Climate-system Historical Forecast Project (CHFP, see Kirtman and Pirani 2009) and the Centro de Investigaciones del Mar y la Atmosfera (CIMA) for providing the model output. We also thank the data providers for making the model output available through CHFP. We have no potential sources of conflict of interest. The contribution of AYK is funded by the Academy of Finland, project #286298.

Supporting Information

Table S1. The area under the relative operative characteristics (ROC) curve (e.g., Doblas-Reyes et al 2003). The columns show model skill for two events: DJF NAO below the lower tercile and above the upper tercile threshold. Statistical significance of the estimated values is tested by the Wilcoxon-Mann-Whitney test. Similarly to Doblas-Reyes et al. (2003), in most cases ROC area is above 0.5 indicating that models possess some skill, although typically it is not statistically significant at p=0.05 (bold values). Also similarly to Doblas-Reyes et al. (2003) we find that the multi-model mean does not perform better than the best individual models.

Figure S1. EOF1 patterns (calculated for 90°W-60°E, 20-90°N SLP anomalies, representing the North Atlantic Oscillation) for NCEP-NCAR reanalysis and individual models. Number in the left corner of each plot is the variance explained by the NAO pattern; number in the right corner of each plot is the spatial correlation to the observed pattern.

Figure S2. Same as Figure 4, but with error bars showing the 95% confidence levels for the correlation with the most ensemble members for each model.

Figure S3. Skill (correlation) of JFM mean sea level pressure anomalies for El Niño winters (as in **Figure 6**), for high-top and low-top pairs. Note that ARPEGE and CMAM do not have March data so the skill is for Jan-Feb averaged anomalies. The values in the upper right of each plot show the number of El Niño winters included in each calculation, and the associated anomaly correlation coefficient (area-weighted correlation over the region shown). Hatching indicates correlations that exceed 95% significance using a 2-tailed t-test.

References

Arribas, A., M. Glover, A. Maidens, K. Peterson, M. Gordon, C. MacLachlan, R. Graham, D. Fereday, J. Camp, A. A. Scaife, P. Xavier, P. McLean, A. Colman, and S. Cusack, 2011: The GloSea4 Ensemble Prediction System for Seasonal Forecasting. *Mon. Weather Rev.*, **139**, 1891–1910, doi:10.1175/2010MWR3615.1.

Baehr, J., K. Fröhlich, M. Botzet, D. I. V Domeisen, L. Kornblueh, D. Notz, R. Piontek, H. Pohlmann, S. Tietsche, and W. A. Müller, 2015: The prediction of surface temperature in the new seasonal prediction system based on the MPI-ESM coupled climate model. *Clim. Dyn.*, **44**, 2723–2735, doi:10.1007/s00382-014-2399-7.

Baldwin, M. P., L. J. Gray, T. J. Dunkerton, K. Hamilton, P. H. Haynes, W. J. Randel, J. R. Holton, M. J. Alexander, I. Hirota, T. Horinouchi, D. B. A. Jones, J. S. Kinnersley, C. Marquardt, K. Sato, and M. Takahashi, 2001: The quasi-biennial oscillation. *Rev. Geophys.*, **39**, 179–229, doi:10.1029/1999RG000073.

Baldwin, M. P., D. B. Stephenson, D. W. J. Thompson, T. J. Dunkerton, A. J. Charlton, A. O'Neill, and David W. J. Thompson, 2003: Stratospheric Memory and Skill of Extended-Range Weather Forecasts. *Science (80-.).*, **301**, 636–640, doi:10.1126/science.1087143.

Becker, E., H. van den Dool, and Q. Zhang, 2014: Predictability and Forecast Skill in NMME. *J. Clim.*, **27**, 5891–5906, doi:10.1175/JCLI-D-13-00597.1.

Boer, G. J., and K. Hamilton, 2008: QBO influence on extratropical predictive skill. *Clim. Dyn.*, **31**, 987–1000, doi:10.1007/s00382-008-0379-5.

Butler, A. H., and L. M. Polvani, 2011: El Niño, La Niña, and stratospheric sudden warmings: A reevaluation in light of the observational record. *Geophys. Res. Lett.*, **38**, doi:10.1029/2011GL048084.

——, ——, and C. Deser, 2014: Separating the stratospheric and tropospheric pathways of El Niño–Southern Oscillation teleconnections. *Environ. Res. Lett.*, **9**, 024014.

Cagnazzo, C., and E. Manzini, 2009: Impact of the Stratosphere on the Winter Tropospheric Teleconnections between ENSO and the North Atlantic and European Region. *J. Clim.*, **22**, 1223–1238, doi:10.1175/2008JCLI2549.1.

Charlton-Perez, A. J., M. P. Baldwin, T. Birner, R. X. Black, A. H. Butler, N. Calvo, N. A. Davis, E. P. Gerber, N. Gillett, S. Hardiman, J. Kim, K. Krüger, Y. Y. Lee, E. Manzini, B. A. McDaniel, L. Polvani, T. Reichler, T. A. Shaw, M. Sigmond, S. W. Son, M. Toohey, L. Wilcox, S. Yoden, B. Christiansen, F. Lott, D. Shindell, S. Yukimoto, and S. Watanabe, 2013: On the lack of stratospheric dynamical variability in low-top versions of the CMIP5 models. *J. Geophys. Res. Atmos.*, **118**, 2494–2505.

Chen, M., W. Wang, and A. Kumar, 2013: Lagged Ensembles, Forecast Configuration, and Seasonal Predictions. *Mon. Weather Rev.*, **141**, 3477–3497, doi:10.1175/MWR-D-12-00184.1.

Cohen, J., and J. Jones, 2011: A new index for more accurate winter predictions. *Geophys. Res. Lett.*, **38**, doi:10.1029/2011GL049626.

Cottrill, A., H. H. Hendon, E.-P. Lim, S. Langford, K. Shelton, A. Charles, D. McClymont, D. Jones, and Y. Kuleshov, 2013: Seasonal Forecasting in the Pacific Using the Coupled Model POAMA-2. *Weather Forecast.*, **28**, 668–680, doi:10.1175/WAF-D-12-00072.1.

Danabasoglu, G., W. G. Large, and B. P. Briegleb, 2010: Climate impacts of parameterized Nordic Sea overflows. *J. Geophys. Res. Ocean.*, **115**, doi:10.1029/2010JC006243.

Dee, D. P., S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut, and F. Vitart, 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.*, **137**, 553–597, doi:10.1002/qj.828.

DelSole, T., J. Nattala, and M. K. Tippett, 2014: Skill Improvement From Increased Ensemble Size and Model Diversity. *Geophys. Res. Lett.*, 2014GL060133, doi:10.1002/2014GL060133.

Doblas-Reyes, F. J., V. Pavan, and D. B. Stephenson, 2003: The skill of multi-model seasonal forecasts of the wintertime North Atlantic Oscillation. *Clim. Dyn.*, **21**, 501–514, doi:10.1007/s00382-003-0350-4.

Domeisen, D. I. V., A. H. Butler, K. Fröhlich, M. Bittner, W. A. Müller, and J. Baehr, 2015: Seasonal Predictability over Europe Arising from El Niño and Stratospheric Variability in the MPI-ESM Seasonal Prediction System. *J. Clim.*, **28**, 256–271, doi:10.1175/JCLI-D-14-00207.1.

Douville, H., 2009: Stratospheric polar vortex influence on Northern Hemisphere winter climate variability. *Geophys. Res. Lett.*, **36**, L18703, doi:10.1029/2009GL039334.

Dunkerton, T., and M. Baldwin, 1991: Quasi-Biennial Modulation of Planetary-Wave Fluxes in the Northern-Hemisphere Winter. *J. Atmos. Sci.*, **48**, 1043–1061, doi:10.1175/1520-0469(1991)048<1043:QBMOPW>2.0.CO;2.

Eade, R., D. Smith, A. Scaife, E. Wallace, N. Dunstone, L. Hermanson, and N. Robinson, 2014: Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophys. Res. Lett.*, **41**, 2014GL061146, doi:10.1002/2014GL061146.

Ern, M., F. Ploeger, P. Preusse, J. C. Gille, L. J. Gray, S. Kalisch, M. G. Mlynczak, J. M. Russell, and M. Riese, 2014: Interaction of gravity waves with the QBO: A satellite perspective. *J. Geophys. Res. Atmos.*, **119**, 2329–2355, doi:10.1002/2013JD020731.

Fereday, D. R., A. Maidens, A. Arribas, A. A. Scaife, and J. R. Knight, 2012: Seasonal forecasts of northern hemisphere winter 2009/10. *Environ. Res. Lett.*, **7**, 034031, doi:10.1088/1748-9326/7/3/034031.

Fletcher, C. G., and P. J. Kushner, 2011: The Role of Linear Interference in the Annular Mode Response to Tropical SST Forcing. *J. Clim.*, **24**, 778–794, doi:10.1175/2010JCLI3735.1.

García-Herrera, R., N. Calvo, R. R. Garcia, and M. a. Giorgetta, 2006: Propagation of ENSO temperature signals into the middle atmosphere: A comparison of two general circulation models and ERA-40 reanalysis data. *J. Geophys. Res.*, **111**, D06101, doi:10.1029/2005JD006061.

García-Serrano, J., C. Frankignoul, G. Gastineau, and Á. de la Cámara, 2015: On the predictability of the winter Euro-Atlantic climate: lagged influence of autumn Arctic sea-ice. *J. Clim.*, doi:10.1175/JCLI-D-14-00472.1.

Garfinkel, C. I., and D. L. Hartmann, 2007: Effects of the El Niño–Southern Oscillation and the Quasi-Biennial Oscillation on polar temperatures in the stratosphere. *J. Geophys. Res.*, **112**, D19112, doi:10.1029/2007JD008481.

——, and ——, 2008: Different ENSO teleconnections and their effects on the stratospheric polar vortex. *J. Geophys. Res.*, **113**, D18114, doi:10.1029/2008JD009920.

——, and ——, 2010: Influence of the quasi-biennial oscillation on the North Pacific and El Niño teleconnections. *J. Geophys. Res.*, **115**, D20116, doi:10.1029/2010JD014181.

——, M. M. Hurwitz, D. W. Waugh, and A. H. Butler, 2013: Are the teleconnections of Central Pacific and Eastern Pacific El Niño distinct in boreal wintertime? *Clim. Dyn.*, 1–18, doi:10.1007/s00382-012-1570-2.

Gerber, E. P., and L. M. Polvani, 2009: Stratosphere–Troposphere Coupling in a Relatively Simple AGCM: The Importance of Stratospheric Variability. *J. Clim.*, **22**, 1920–1933, doi:10.1175/2008JCLI2548.1.

Gerber, E. P., C. Orbe, and L. M. Polvani, 2009: Stratospheric influence on the tropospheric circulation revealed by idealized ensemble forecasts. *Geophys. Res. Lett.*, **36**, L24801, doi:10.1029/2009GL040913.

Gerber, E. P., A. Butler, N. Calvo, A. Charlton-Perez, M. Giorgetta, E. Manzini, J. Perlwitz, L. M. Polvani, F. Sassi, A. A. Scaife, T. A. Shaw, S.-W. Son, and S. Watanabe, 2012: Assessing and Understanding the Impact of Stratospheric Dynamics and Variability on the Earth System. *Bull. Am. Meteorol. Soc.*, **93**, 845–859, doi:10.1175/BAMS-D-11-00145.1.

Hardiman, S. C., N. Butchart, T. J. Hinton, S. M. Osprey, and L. J. Gray, 2012: The Effect of a Well-Resolved Stratosphere on Surface Climate: Differences between CMIP5 Simulations with High and Low Top Versions of the Met Office Climate Model. *J. Clim.*, **25**, 7083–7099, doi:10.1175/JCLI-D-11-00579.1.

Haynes, P. H., 1998: The latitudinal structure of the quasi-biennial oscillation. *Q. J. R. Meteorol. Soc.*, **124**, 2645–2670, doi:10.1002/qj.49712455206.

Holton, J. R., and H.-C. Tan, 1980: The Influence of the Equatorial Quasi-Biennial Oscillation on the Global Circulation at 50 mb. *J. Atmos. Sci.*, **37**, 2200–2208, doi:10.1175/1520-0469(1980)037<2200:TIOTEQ>2.0.CO;2.

Hurwitz, M. M., N. Calvo, C. I. Garfinkel, A. H. Butler, S. Ineson, C. Cagnazzo, E. Manzini, and C. Peña-Ortiz, 2014: Extra-tropical atmospheric response to ENSO in the CMIP5 models. *Clim. Dyn.*, doi:10.1007/s00382-014-2110-z.

Imada, Y., H. Tatebe, M. Ishii, Y. Chikamoto, M. Mori, M. Arai, M. Watanabe, and M. Kimoto, 2015: Predictability of Two Types of El Niño Assessed Using an Extended Seasonal Prediction System by MIROC. *Mon. Weather Rev.*, doi:10.1175/MWR-D-15-0007.1.

Ineson, S., and A. A. Scaife, 2009: The role of the stratosphere in the European climate response to El Niño. *Nat. Geosci.*, **2**, 32–36, doi:10.1038/ngeo381.

Iza, M., and N. Calvo, 2015: Role of Stratospheric Sudden Warmings on the response to Central Pacific El Niño. *Geophys. Res. Lett.*, **42**, 2482–2489, doi:10.1002/2014GL062935.

Jungclaus, J. H., N. Fischer, H. Haak, K. Lohmann, J. Marotzke, D. Matei, U. Mikolajewicz, D. Notz, and J. S. von Storch, 2013: Characteristics of the ocean simulations in the Max Planck Institute Ocean Model (MPIOM) the ocean component of the MPI-Earth system model. *J. Adv. Model. Earth Syst.*, **5**, 422–446, doi:10.1002/jame.20023.

Kang, D., M.-I. Lee, J. Im, D. Kim, H.-M. Kim, H.-S. Kang, S. D. Schubert, A. Arribas, and C. MacLachlan, 2014: Prediction of the Arctic Oscillation in boreal winter by dynamical seasonal forecasting systems. *Geophys. Res. Lett.*, **41**, 2014GL060011, doi:10.1002/2014GL060011.

Keeley, S. P. E., R. T. Sutton, and L. C. Shaffrey, 2009: Does the North Atlantic Oscillation show unusual persistence on intraseasonal timescales? *Geophys. Res. Lett.*, **36**, doi:10.1029/2009GL040367.

Kharin, V. V, F. W. Zwiers, and N. Gagnon, 2001: Skill of seasonal hindcasts as a function of the ensemble size. *Clim. Dyn.*, **17**, 835–843, doi:10.1007/s003820100149.

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

Kim, H.-M., P. J. Webster, and J. A. Curry, 2012: Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere Winter. *Clim. Dyn.*, **39**, 2957–2973, doi:10.1007/s00382-012-1364-6.

Kirtman, B., and A. Pirani, 2009: The State of the Art of Seasonal Prediction: Outcomes and Recommendations from the First World Climate Research Program Workshop on Seasonal Prediction. *Bull. Am. Meteorol. Soc.*, **90**, 455–458, doi:10.1175/2008BAMS2707.1.

Kumar, A., 2009: Finite Samples and Uncertainty Estimates for Skill Measures for Seasonal Prediction. *Mon. Weather Rev.*, **137**, 2622–2631, doi:10.1175/2009MWR2814.1.

——, and M. P. Hoerling, 2000: Analysis of a Conceptual Model of Seasonal Climate Variability and Implications for Seasonal Prediction. *Bull. Am. Meteorol. Soc.*, **81**, 255–264, doi:10.1175/1520-0477(2000)081<0255:AOACMO>2.3.CO;2.

Kuroda, Y., 2008: Role of the stratosphere on the predictability of medium-range weather forecast: A case study of winter 2003–2004. *Geophys. Res. Lett.*, **35**, L19701, doi:10.1029/2008GL034902.

Lee, T., and M. J. McPhaden, 2010: Increasing intensity of El Niño in the central-equatorial Pacific. *Geophys. Res. Lett.*, **37**, L14603, doi:10.1029/2010GL044007.

MacLachlan, C., A. Arribas, K. A. Peterson, A. Maidens, D. Fereday, A. A. Scaife, M. Gordon, M. Vellinga, A. Williams, R. E. Comer, J. Camp, P. Xavier, and G. Madec, 2015: Global Seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system. *Q. J. R. Meteorol. Soc.*, **141**, 1072–1084, doi:10.1002/qj.2396.

Manzini, E., M. a. Giorgetta, M. Esch, L. Kornblueh, and E. Roeckner, 2006: The Influence of Sea Surface Temperatures on the Northern Winter Stratosphere: Ensemble Simulations with the MAECHAM5 Model. *J. Clim.*, **19**, 3863–3881, doi:10.1175/JCLI3826.1.

Marshall, A. G., and A. a. Scaife, 2009: Impact of the QBO on surface winter climate. *J. Geophys. Res.*, **114**, D18110, doi:10.1029/2009JD011737.

——, and ——, 2010: Improved predictability of stratospheric sudden warming events in an atmospheric general circulation model with enhanced stratospheric resolution. *J. Geophys. Res.*, **115**, D16114, doi:10.1029/2009JD012643.

Maycock, A. C., S. P. E. Keeley, A. J. Charlton-Perez, and F. J. Doblas-Reyes, 2011: Stratospheric circulation in seasonal forecasting models: implications for seasonal prediction. *Clim. Dyn.*, **36**, 309–321, doi:10.1007/s00382-009-0665-x.

Merryfield, W. J., W.-S. Lee, G. J. Boer, V. V Kharin, J. F. Scinocca, G. M. Flato, R. S. Ajayamohan, J. C. Fyfe, Y. Tang, and S. Polavarapu, 2013: The Canadian Seasonal to Interannual Prediction System. Part I: Models and Initialization. *Mon. Weather Rev.*, **141**, 2910–2945, doi:10.1175/MWR-D-12-00216.1.

Molteni, F., T. Stockdale, M. Balmaseda, G. Balsamo, R. Buizza, L. Ferranti, L. Magnusson, K. Mogensen, T. Palmer, and F. Vitart, 2011: The new ECMWF seasonal forecast system (System 4). *Tech. Memo.*, **656**, 1–49.

Müller, W. A., C. Appenzeller, and C. Schär, 2005: Probabilistic seasonal prediction of the winter North Atlantic Oscillation and its impact on near surface temperature. *Clim. Dyn.*, **24**, 213–226, doi:10.1007/s00382-004-0492-z.

National Research Council, 2010: *Assessment of Intraseasonal to Interannual Climate Prediction and Predictability*. The National Academies Press, Washington, DC,.

Orsolini, Y. J., I. T. Kindem, and N. G. Kvamstø, 2009: On the potential impact of the stratosphere upon seasonal dynamical hindcasts of the North Atlantic Oscillation: a pilot study. *Clim. Dyn.*, **36**, 579–588, doi:10.1007/s00382-009-0705-6.

Osprey, S. M., L. J. Gray, S. C. Hardiman, N. Butchart, and T. J. Hinton, 2013: Stratospheric Variability in Twentieth-Century CMIP5 Simulations of the Met Office Climate Model: High Top versus Low Top. *J. Clim.*, **26**, 1595–1606, doi:10.1175/JCLI-D-12-00147.1.

Peng, P., A. Kumar, H. van den Dool, and A. G. Barnston, 2002: An analysis of multimodel ensemble predictions for seasonal climate anomalies. *J. Geophys. Res. Atmos.*, **107**, 4710, doi:10.1029/2002JD002712.

Riddle, E. E., A. H. Butler, J. C. Furtado, J. L. Cohen, and A. Kumar, 2013: CFSv2 ensemble prediction of the wintertime Arctic Oscillation. *Clim. Dyn.*, **41**, 1099–1116.

Roff, G., D. W. J. Thompson, and H. Hendon, 2011: Does increasing model stratospheric resolution improve extended-range forecast skill? *Geophys. Res. Lett.*, **38**, doi:10.1029/2010GL046515.

Saha, S., S. Nadiga, C. Thiaw, J. Wang, W. Wang, Q. Zhang, H. M. Van den Dool, H.-L. Pan, S. Moorthi, D. Behringer, D. Stokes, M. Peña, S. Lord, G. White, W. Ebisuzaki, P. Peng, and P. Xie, 2006: The NCEP Climate Forecast System. *J. Clim.*, **19**, 3483–3517, doi:10.1175/JCLI3812.1.

Von Salzen, K., J. F. Scinocca, N. A. McFarlane, J. Li, J. N. S. Cole, D. Plummer, D. Verseghy, M. C. Reader, X. Ma, M. Lazare, and L. Solheim, 2013: The Canadian Fourth Generation Atmospheric Global Climate Model (CanAM4). Part I: Representation of Physical Processes. *Atmosphere-Ocean*, **51**, 104–125, doi:10.1080/07055900.2012.755610.

Sato, K., and T. J. Dunkerton, 1997: Estimates of momentum flux associated with equatorial Kelvin and gravity waves. *J. Geophys. Res. Atmos.*, **102**, 26247–26261, doi:10.1029/96JD02514.

Scaife, A. A., J. R. Knight, G. K. Vallis, and C. K. Folland, 2005: A stratospheric influence on the winter NAO and North Atlantic surface climate. *Geophys. Res. Lett.*, **32**, doi:10.1029/2005GL023226.

——, D. Copsey, C. Gordon, C. Harris, T. Hinton, S. Keeley, A. O’Neill, M. Roberts, and K. Williams, 2011: Improved Atlantic winter blocking in a climate model. *Geophys. Res. Lett.*, **38**, doi:10.1029/2011GL049573.

——, T. Spangehl, D. R. Fereday, U. Cubasch, U. Langematz, H. Akiyoshi, S. Bekki, P. Braesicke, N. Butchart, M. P. Chipperfield, A. Gettelman, S. C. Hardiman, M. Michou, E. Rozanov, and T. G. Shepherd, 2012: Climate change projections and stratosphere–troposphere interaction. *Clim. Dyn.*, **38**, 2089–2097, doi:10.1007/s00382-011-1080-7.

Scaife, A. A., A. Arribas, E. Blockley, A. Brookshaw, R. T. Clark, N. Dunstone, R. Eade, D. Fereday, C. K. Folland, M. Gordon, L. Hermanson, J. R. Knight, D. J. Lea, C. MacLachlan, A. Maidens, M. Martin, A. K. Peterson, D. Smith, M. Vellinga, E. Wallace, J. Waters, and A. Williams, 2014a: Skillful long-range prediction of European and North American winters. *Geophys. Res. Lett.*, **41**, 2014GL059637, doi:10.1002/2014GL059637.

Scaife, A. A., M. Athanassiadou, M. Andrews, A. Arribas, M. Baldwin, N. Dunstone, J. Knight, C. MacLachlan, E. Manzini, W. A. Müller, H. Pohlmann, D. Smith, T.

Stockdale, and A. Williams, 2014b: Predictability of the Quasi-Biennial Oscillation and its Northern Winter Teleconnection on Seasonal to Decadal Timescales. *Geophys. Res. Lett.*, **41**, 1752–1758, doi:10.1002/2013GL059160.

Scaife, A. A., A. Y. Karpechko, M. P. Baldwin, A. Brookshaw, A. H. Butler, R. Eade, M. Gordon, C. MacLachlan, N. Martin, N. Dunstone, and D. Smith, 2015: Seasonal winter forecasts and the stratosphere. *Atmos. Sci. Lett.*, doi:10.1002/asl.598.

Scinocca, J. F., N. A. McFarlane, M. Lazare, J. Li, and D. Plummer, 2008: Technical Note: The CCCma third generation AGCM and its extension into the middle atmosphere. *Atmos. Chem. Phys.*, **8**, 7055–7074, doi:10.5194/acp-8-7055-2008.

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

——, ——, and O. Weiner, 2014: Troposphere-stratosphere coupling: Links to North Atlantic weather and climate, including their representation in CMIP5 models. *J. Geophys. Res. Atmos.*, 2013JD021191, doi:10.1002/2013JD021191.

Shi, W., N. Schaller, D. MacLeod, T. N. Palmer, and A. Weisheimer, 2015: Impact of hindcast length on estimates of seasonal climate predictability. *Geophys. Res. Lett.*, **42**, 1554–1559, doi:10.1002/2014GL062829.

Sigmond, M., J. F. Scinocca, and P. J. Kushner, 2008: Impact of the stratosphere on tropospheric climate change. *Geophys. Res. Lett.*, **35**, doi:10.1029/2008GL033573.

Sigmond, M., J. F. Scinocca, V. V. Kharin, and T. G. Shepherd, 2013: Enhanced seasonal forecast skill following stratospheric sudden warmings. *Nat. Geosci.*, **6**, 98–102, doi:10.1038/ngeo1698.

Smith, D. M., A. A. Scaife, and B. P. Kirtman, 2012: What is the current state of scientific knowledge with regard to seasonal and decadal forecasting? *Environ. Res. Lett.*, **7**, 15602.

Smith, K. L., C. G. Fletcher, and P. J. Kushner, 2010: The Role of Linear Interference in the Annular Mode Response to Extratropical Surface Forcing. *J. Clim.*, **23**, 6036–6050, doi:10.1175/2010JCLI3606.1.

Stevens, B., M. Giorgetta, M. Esch, T. Mauritsen, T. Crueger, S. Rast, M. Salzmann, H. Schmidt, J. Bader, K. Block, R. Brokopf, I. Fast, S. Kinne, L. Kornblueh, U. Lohmann, R. Pincus, T. Reichler, and E. Roeckner, 2013: Atmospheric component of the MPI-M Earth System Model: ECHAM6. *J. Adv. Model. Earth Syst.*, **5**, 146–172, doi:10.1002/jame.20015.

Stockdale, T. N., F. Molteni, and L. Ferranti, 2015: Atmospheric initial conditions and the predictability of the Arctic Oscillation. *Geophys. Res. Lett.*, **42**, 2014GL062681, doi:10.1002/2014GL062681.

Takaya, Y., T. Yasuda, T. Ose, and T. Nakaegawa, 2010: Predictability of the Mean Location of Typhoon Formation in a Seasonal Prediction Experiment with a Coupled General Circulation Model. *J. Meteorol. Soc. Japan*, **88**, 799–812, doi:10.2151/jmsj.2010-502.

Toniazzo, T., and A. a. Scaife, 2006: The influence of ENSO on winter North Atlantic climate. *Geophys. Res. Lett.*, **33**, L24704, doi:10.1029/2006GL027881.

Tripathi, O. P., M. Baldwin, A. Charlton-Perez, M. Charron, S. D. Eckermann, E. Gerber, R. G. Harrison, D. R. Jackson, B.-M. Kim, Y. Kuroda, A. Lang, S. Mahmood, R. Mizuta, G. Roff, M. Sigmond, and S.-W. Son, 2014: The predictability of the extratropical stratosphere on monthly time-scales and its impact on the skill of tropospheric forecasts. *Q. J. R. Meteorol. Soc.*, doi:10.1002/qj.2432.

——, A. Charlton-Perez, M. Sigmond, and F. Vitart, 2015: Enhanced long-range forecast skill in boreal winter following stratospheric strong vortex conditions. *Environ. Res. Lett.*, **10**, 104007.

Voldoire, A., E. Sanchez-Gomez, D. Salas y Mélia, B. Decharme, C. Cassou, S. Sénési, S. Valcke, I. Beau, A. Alias, M. Chevallier, M. Déqué, J. Deshayes, H. Douville, E. Fernandez, G. Madec, E. Maisonnave, M.-P. Moine, S. Planton, D. Saint-Martin, S. Szopa, S. Tyteca, R. Alkama, S. Belamari, A. Braun, L. Coquart, and F. Chauvin, 2013: The CNRM-CM5.1 global climate model: description and basic evaluation. *Clim. Dyn.*, **40**, 2091–2121, doi:10.1007/s00382-011-1259-y.

Watanabe, M., T. Suzuki, R. O'ishi, Y. Komuro, S. Watanabe, S. Emori, T. Takemura, M. Chikira, T. Ogura, M. Sekiguchi, K. Takata, D. Yamazaki, T. Yokohata, T. Nozawa, H. Hasumi, H. Tatebe, and M. Kimoto, 2010: Improved Climate Simulation by MIROC5: Mean States, Variability, and Climate Sensitivity. *J. Clim.*, **23**, 6312–6335, doi:10.1175/2010JCLI3679.1.

Table I. List of models participating in CHFP. The high-top models are in bolded italics.

Model	# of ens mem	Horiz Res	Vert Res	Approx Model lid height (hPa)	Time period (Nov-Mar)	Months forecast (Nov-Mar)	Output levels above sfc (hPa)	Reference	Initial Atmosphere Ocean Configuration Sea Ice Configuration
ARPEGE_z00k	11	T63	L91	0.01	79-08	NDJF	1000, 925, 850, 700, 500, 400, 300, 200, 50, 30, 10	Voldoire et al 2013	ERA40/ERA-Interim Interactive Ocean Climatological Sea Ice
ARPEGE_z00l	11	T63	L31	10	79-08	NDJF	1000, 925, 850, 700, 500, 400, 300, 200, 50, 30, 10	Voldoire et al 2013	ERA40/ERA-Interim Interactive Ocean Climatological Sea Ice
CCCma-CanCM3	10	T63	L31	1	79-11	NDJFM	850, 500, 200, 100, 50	Scinocca et al 2008; Merryfield et al 2013	ERA-40/Era-Interim Interactive Ocean Interactive Sea Ice
CCCma-CanCM4	10	T63	L35	1	79-11	NDJFM	850, 500, 200, 100, 50	Merryfield et al 2013; von Salzen et al 2013	ERA-40/ERA-Interim Interactive Ocean Interactive Sea Ice
CFSv1	7	T62	L64	0.2	81-07	NDJFM	1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, 70, 50, 30, 20, 10	Saha et al 2006	NCEP DOE Interactive Ocean Climatological Sea ice
CMAM	10	T63	L71	0.0005	79-09	NDJF	1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, 70, 50, 30, 20, 10, 7, 5, 3, 2, 1, .5, .4	Scinocca et al 2008	ERA-40/ERA-Interim Persisted SSTA Climatological Sea ice

CMAMlo	10	T63	L41	10	79-09	NDJF	1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, 70, 50, 30, 20, 10	Sigmond et al 2008	ERA-40/ERA-Interim Persisted SSTA Climatological Sea ice
ECMWF-S4	15	T255	L91	0.01	81-11	NDJFM	850, 500, 200, 100, 50, 10	Molteni et al 2011	ERA-Interim Interactive Ocean Sea Ice sampled from previous 5 years
GloSea5	24	N216	L85	0.005	96-10	DJF	850, 200, 50, 10 (ua); 850, 500, 200 (z)	MacLachlan et al 2015	ERA-Interim Interactive Ocean Interactive Sea Ice
JMA/MRI-CGCM1	10	T95	L40	0.4	79-10	NDJFM	850, 500, 200	Takaya et al 2010	JRA-25/JCDAS Interactive Ocean Climatological Sea Ice
L85GloSea4	9	N96	L85	0.005	89-10	DJFM	850, 200	Fereday et al 2012	ERA-Interim Interactive Ocean Interactive Sea Ice
L38GloSea4	9	N96	L38	3	89-03	DJFM	850, 200	Arribas et al 2011	ERA-Interim Interactive Ocean Interactive Sea Ice
MIROC5	8	T85	L40	3	79-12	DJFM	850, 500, 200, 100, 50, 10	Watanabe et al 2010; Imada et al 2015	NCEP/NCAR 1 Interactive Ocean Interactive Sea Ice
MPI-ESM-LR	9	T63	L47	0.01	82-12	NDJFM	850, 500, 200, 100, 50, 10	Baehr et al 2015	ERA-Interim Interactive ocean Interactive Sea Ice
MPI-ESM-MR	10	T63	L95	0.01	81-12	NDJFM	850, 500, 200, 100, 50, 10	Stevens et al. 2013; Jungclaus et al. 2013	ERA-Interim Interactive ocean Interactive Sea Ice
POAMA 2.4	10	T47	L17	10	80-10	DJFM	850, 500, 200	Cottrill et al	ERA-Interim through

(p24a, b, c)	each; 30 total							2013	2002, POAMA NWP global analyses after Interactive ocean Climatological Sea Ice

Table II. El Niño, La Niña, ENSO-neutral, easterly QBO, and westerly QBO years, as defined in Section 2. The number of each phase is given in parentheses.

Winter	El Niño (10)	La Niña (12)	EQBO (12)	WQBO (14)
1979-80			X	
1980-81				X
1981-82			X	
1982-83	X			X
1983-84		X		
1984-85		X	X	
1985-86				X
1986-87	X			
1987-88	X			X
1988-89		X		X
1989-90			X	
1990-91				X
1991-92	X		X	
1992-93				
1993-94				
1994-95	X		X	
1995-96		X		
1996-97			X	
1997-98	X			X
1998-99		X	X	
1999-00		X		X
2000-01		X		
2001-02			X	
2002-03	X			X
2003-04			X	
2004-05	X			X
2005-06		X	X	
2006-07	X			X
2007-08		X	X	
2008-09		X		X
2009-10	X			
2010-11		X		X
2011-12		X		X

Table III. Correlation of model-mean DJF-mean NAO and NAM indexes with ERA-Interim. The NAO is defined as the 1st EOF of monthly mean sea level pressure (SLP) anomalies over the region 90°W-60°E and 20°N-90°N; or as the difference between the normalized SLP times series between the Azores Islands (38°N, 26°W) and Iceland (64°N, 22°W). The NAM is defined as the 1st EOF of zonal mean monthly mean 500-hPa geopotential height anomalies 20°N-90°N. Patterns were calculated based on detrended, deseasonalized DJF monthly anomalies using all ensemble members of each model. Indexes were calculated by projecting anomalies onto model's own pattern. Numbers in parentheses indicate significance level (the smaller the better). Numbers in bold indicate coefficients significant at *p*<0.05.

Model name	Period (Nov Yr 1 - Mar Yr end)	N yrs	Ens size	NAOi skill			NAM-500 hPa skill
				EOF	Azores-Iceland	1996 to ≤ 2009	
ARPEGE_z00k	1979-2008	29	11	0.28 (0.14)	0.21 (0.27)	-0.05 (0.88)	0.28 (0.15)
CFS	1981-2007	26	7	0.14 (0.48)	0.12 (0.56)	-0.11 (0.75)	0.03 (0.90)
CMAM	1979-2009	30	10	0.29 (0.12)	0.10 (0.60)	0.13 (0.67)	0.12 (0.53)
ECMWF-S4	1981-2011	30	15	0.26 (0.16)	0.05 (0.79)	-0.25 (0.39)	0.32 (0.08)
GloSea5	1996-2010	14	24	**0.63** **(0.02)**	**0.61** **(0.02)**	**0.61** **(0.02)**	**0.67** **(0.01)**
L85GloSea4	1989-2010	21	9	0.07 (0.77)	0.09 (0.70)	0.04 (0.89)	0.26 (0.26)
MPI-ESM-LR	1982-2012	30	9	0.13 (0.50)	0.18 (0.34)	0.27 (0.35)	0.27 (0.15)
MPI-ESM-MR	1981-2012	31	10	**0.38** **(0.03)**	**0.43** **(0.02)**	**0.71** **(0.004)**	0.32 (0.08)
High-top EM	1979-2012	33	95	**0.45** **(0.01)**	**0.42** **(0.02)**	0.37 (0.19)	0.32 (0.07)
ARPEGE_z00l	1979-2008	29	11	**0.43** **(0.02)**	0.20 (0.29)	0.06 (0.85)	0.20 (0.29)
CCCma-CanCM3	1979-2011	32	10	0.16 (0.38)	0.24 (0.19)	0.15 (0.61)	0.14 (0.44)
CCCma-CanCM4	1979-2011	32	10	0.31 (0.08)	0.25 (0.17)	0.37 (0.19)	**0.40** **(0.02)**
CMAMlo	1979-2009	30	10	0.17 (0.38)	0.25 (0.18)	0.45 (0.12)	0.07 (0.73)
JMA/MRI-CGCM1	1979-2011	32	10	-0.01 (0.94)	0.23 (0.21)	0.39 (0.17)	-0.02 (0.90)
L38GloSea4	1989-2003	14	9	0.24 (0.41)	0.22 (0.45)	0.68 (0.09)	0.34 (0.23)
MIROC5	1979-2012	33	8	0.07 (0.70)	0.30 (0.09)	0.12 (0.68)	0.30 (0.09)

poama	1980-2010	30	30	-0.01 (0.95)	0.03 (0.87)	0.21 (0.47)	-
Low-top EM	1979-2012	33	98/ 68*	0.32 (0.07)	**0.44 (0.01)**	0.44 (0.12)	0.31 (0.08)

(*) Low top ensemble has 98 members for NAO and 68 members for NAM.

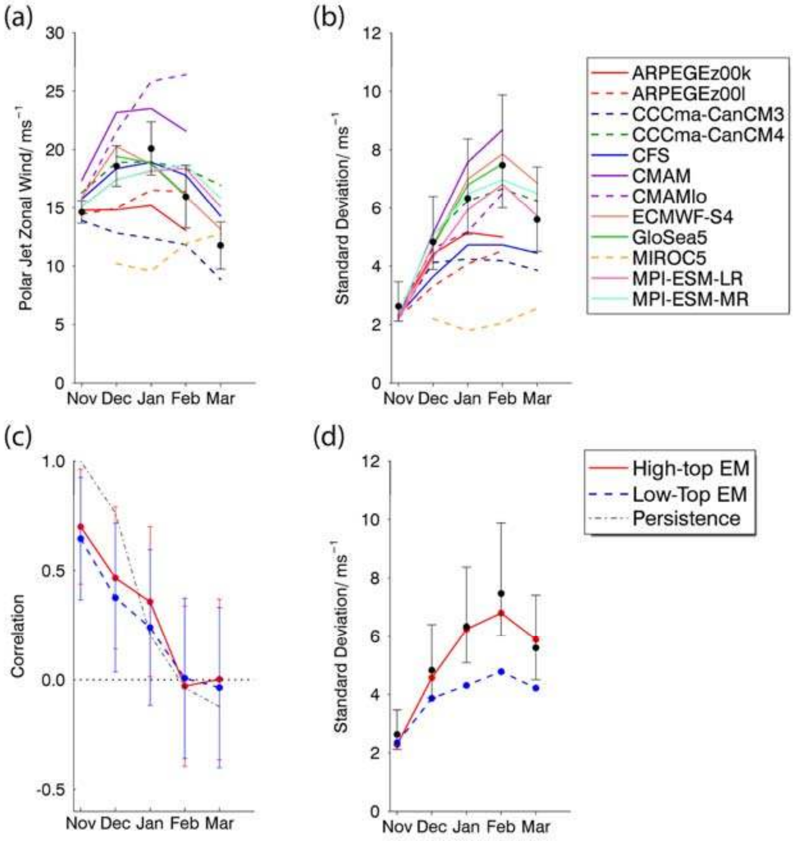

Figure 1. Individual model performance of (a) NH wintertime stratospheric jet (zonal winds area-weighted from 50-70°N at 50 hPa) and (b) its standard deviation from November-March. High-top models are denoted with a solid line; low-top models are denoted with a dashed line. Black dots show results from ERA-interim reanalysis for winters from Nov 1979- Mar 2012. (c) Skill of multi-model mean for the NH wintertime stratospheric jet, and (d) average standard deviation. High-top ensemble mean is the solid line, low-top ensemble mean is the dashed line, and in (c), the "persistence" forecast is the thin dashed-dot line. The black error bars in (a) indicate the 95% confidence interval for the observed mean using a 2-tailed *t*-test. For the standard deviations in (b) and (d), the 95% confidence interval using the chi-squared distribution is shown. In (c), the error bars indicate a 95% confidence interval for the ensemble-mean correlation coefficient using a 2-tailed *t*-test.

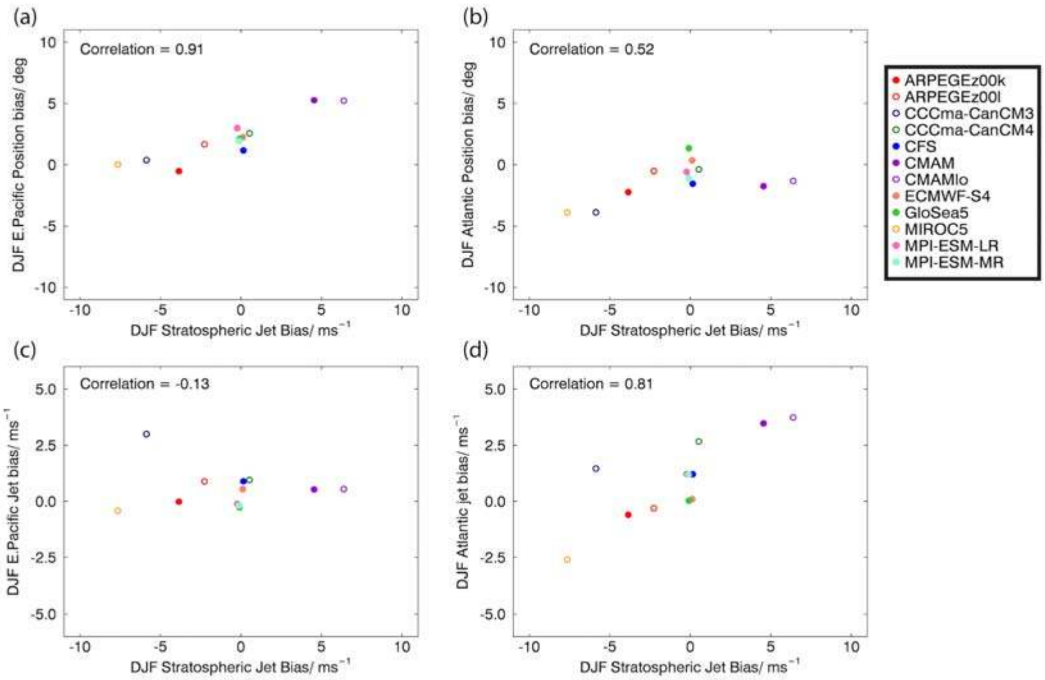

Figure 2. Scatterplots of the bias in the strength of the DJF-mean, 50-70°N, 50 hPa zonal winds (i.e., the NH stratospheric polar jet) versus the bias in the *position* of the DJF-mean tropospheric jet at 850 hPa in the (a) East Pacific and (b) Atlantic; and versus the DJF-mean tropospheric jet *strength* at 850 hPa and 35-55°N in the (c) East Pacific and (d) Atlantic. In all cases, biases are calculated relative to the ERA-interim climatology for the period Nov 1979- Mar 2012. High-top models are shown with a solid dot and low-top models with an empty circle. The correlation between the biases or the skill for all models is shown in the top left corner of each plot. The correlations in panel (a) and (d) are significant at $p<0.01$ using the *2*-tailed *t*-test.

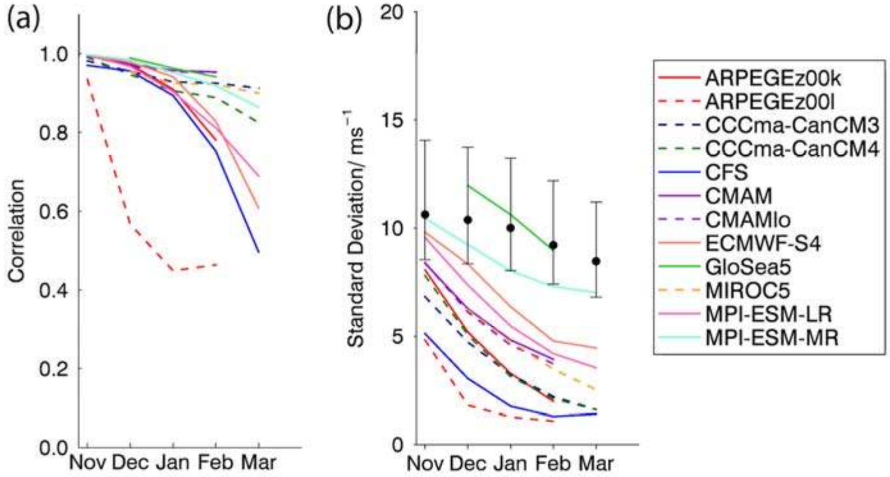

Figure 3. Individual model diagnostics for stratospheric tropical winds (10°S-10°N, 50 hPa) and ERA-interim (black dots) from November-March. (left) Correlation skill for each model. High-top models are denoted with a solid line; low-top models are denoted with a dashed line. (right) Standard deviation [m s^{-1}] of tropical winds in each model. Error bars for ERA-interim (black dots) are given by the 95% confidence interval using the chi-squared distribution.

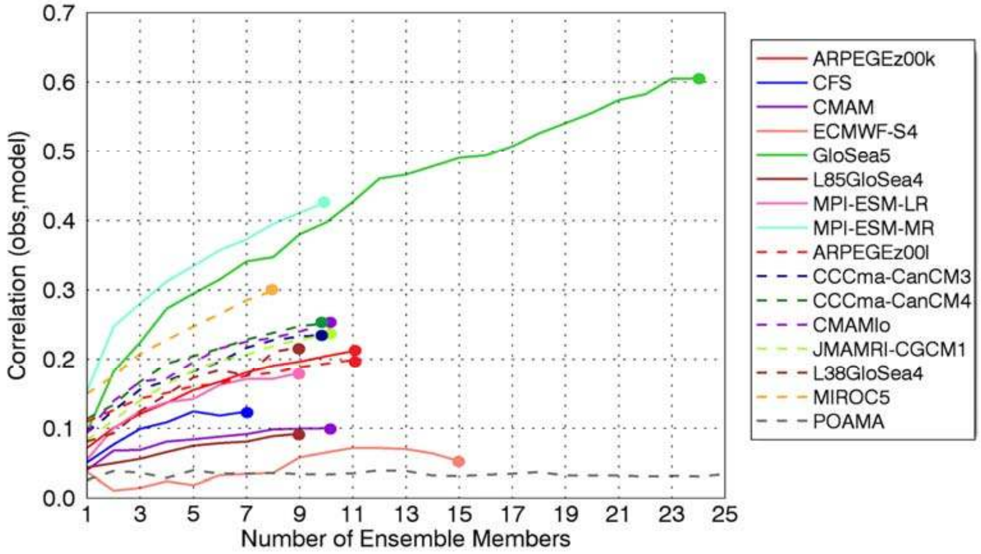

Figure 4. Skill of model forecasts of the DJF NAO index (here calculated as the SLP difference between the Icelandic low and Azores high pressure centers) versus the number of ensemble members. Here the correlations are calculated by randomly selecting ensemble members 100 times, averaging them together if the number of ensemble members is greater than 1, and then correlating the ensemble-mean with the observed DJF NAO index using ERA-interim reanalysis.

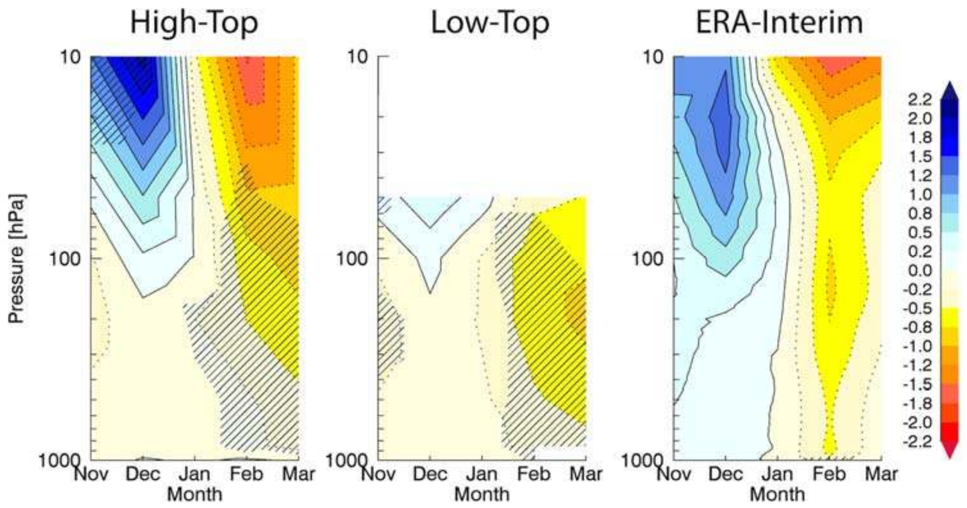

Figure 5. The El Niño zonal wind anomaly [m s^{-1}] response averaged 50-80°N from November to March, for the high-top models (left), the low-top models (center), and ERA-interim (right). For each model, the ensemble-mean climatology for each month is removed to create anomalies. El Niño/La Niña winters are composited and then averaged over ensemble members to get the ensemble-mean response for that model. The *t*-test for differences between two means is used to determine significance at the 95% level for each model-mean. The response is shown only at selected common pressure levels (1000, 850, 500, 200, 100, 50, 10 hPa). If the model has no data at those levels, missing data is used. Here, the significance for the high-top and low-top plots (indicated by hatching) is determined by where at least half of the models (as a function of the total models that have valid data at a given gridpoint) show significant differences between El Niño and La Niña at the 95% level.

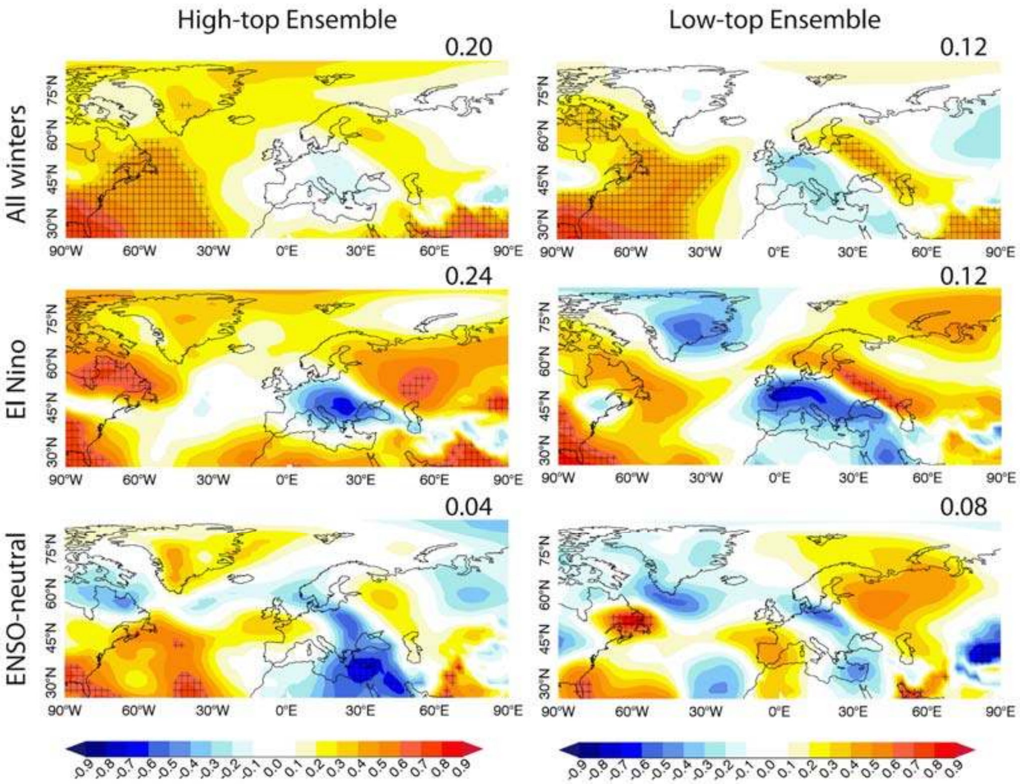

Figure 6. The skill (correlation) of JFM mean sea level pressure anomalies at each grid cell from 1980-2012 for the high-top ensemble (left panels) and low-top ensemble (right panels) for all winters (top row), El Niño winters only (middle row), and ENSO-neutral winters (bottom row). The anomaly correlation coefficient (ACC) for the region shown, per equation 4, is given in the upper right corner of each plot. Hatching indicates correlations that exceed 95% significance using a 2-tailed t-test.

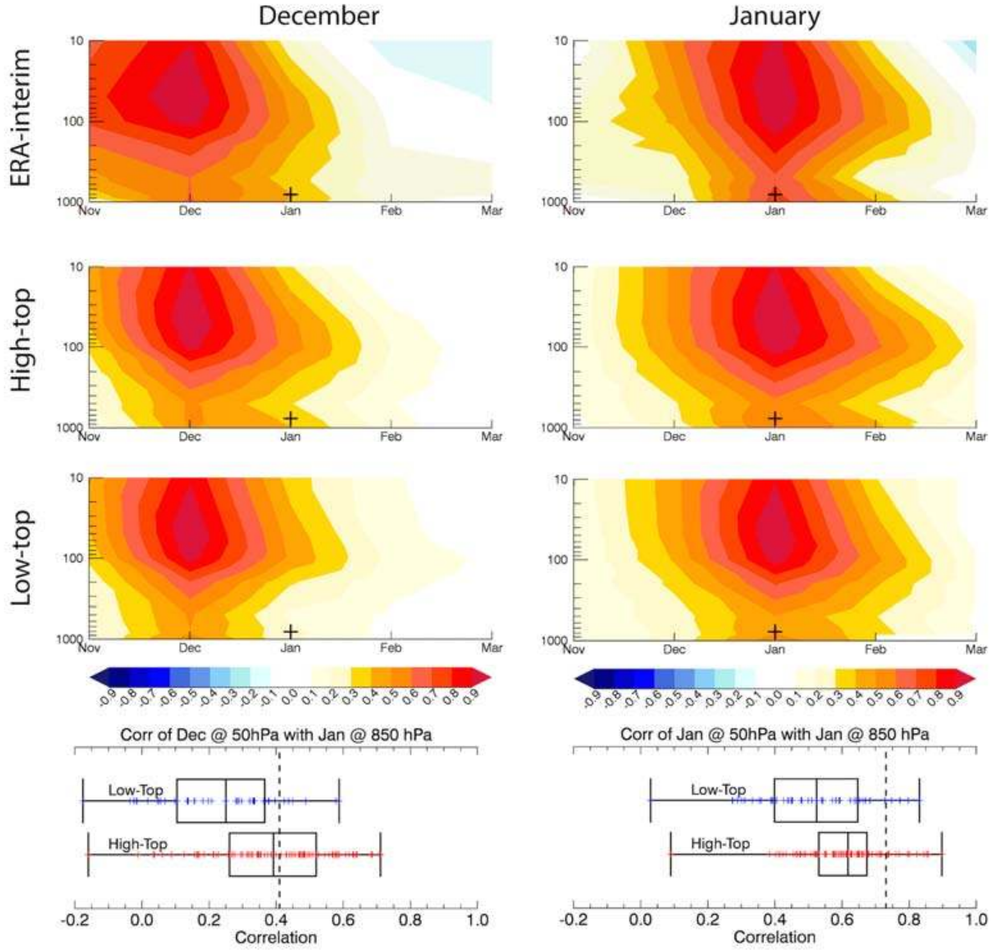

Figure 7. Correlation of 50 hPa, 50-80°N zonal wind anomalies in December (left column) and January (right column) with zonal wind anomalies at every other pressure level and month, for ERA-interim (top), high-top models (second row), and low-top models (third row). For model data, correlations are found for individual model ensemble members and then averaged together for each model, then averaged into high-top or low-top ensembles (so each model is weighted equally). Bottom row shows a box and whisker plot of the correlations (at the location marked at the cross in the panels above) for the ensemble members of the low-top and high-top models, with the dashed line showing the observed correlation.

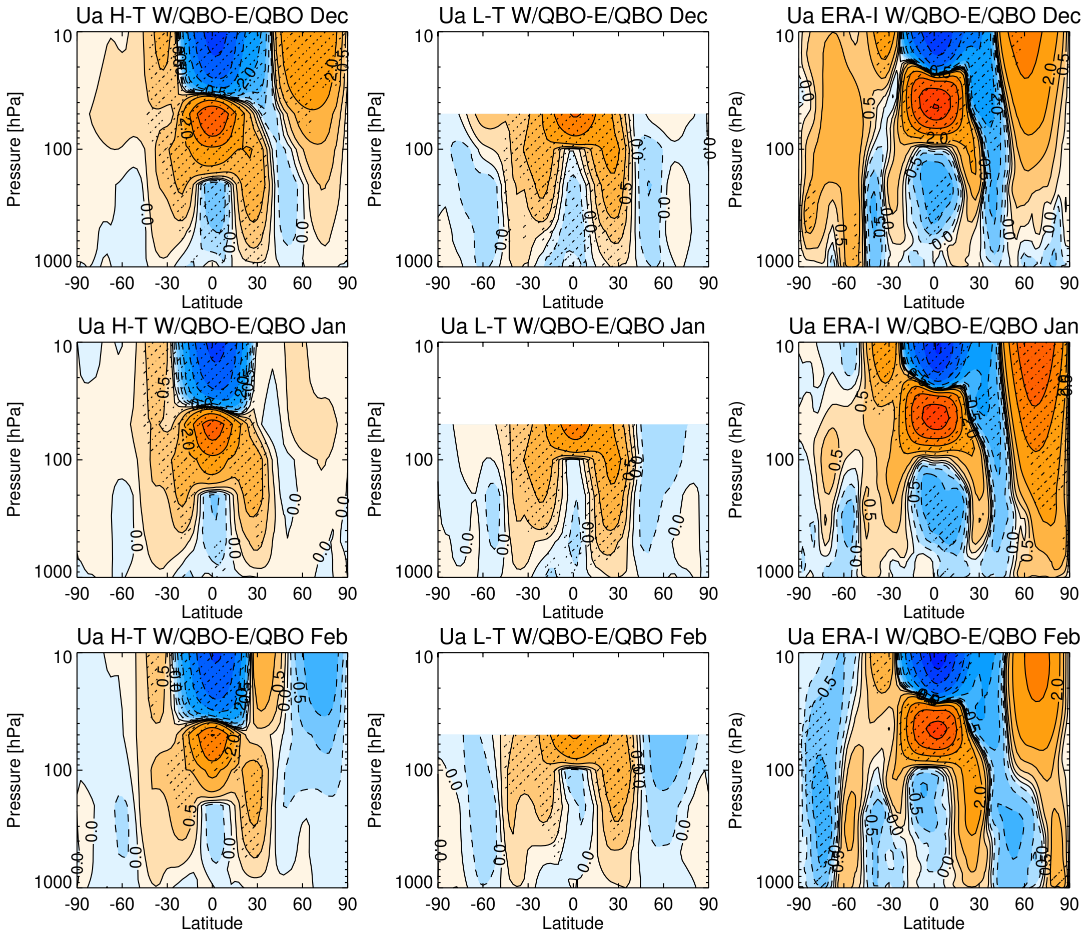

Figure 8. Westerly minus easterly QBO composites of zonal wind anomalies as a function of latitude vs. pressure, for (top row) December, (middle row) January, and (bottom row) February. (Left) Composite for high-top models, (middle) composite for low-top models. Composites are created in the same manner as Figure 5. Significance for the high-top and low-top figures (indicated by shading) is determined by where at least half of the models are significant at the 95% level. (Right) ERA-Interim, shadowed regions are significant at the 95% level according to a Monte Carlo test.