Clim Dyn (2013) 41:1099–1116
DOI 10.1007/s00382-013-1850-5

CFSv2 ensemble prediction of the wintertime Arctic Oscillation

Emily E. Riddle · Amy H. Butler · Jason C. Furtado · Judah L. Cohen · Arun Kumar

Received: 27 February 2013 / Accepted: 17 June 2013 / Published online: 2 July 2013

Abstract Lagged ensembles from the operational Climate Forecast System version 2 (CFSv2) seasonal hindcast dataset are used to assess skill in forecasting interannual variability of the December–February Arctic Oscillation (AO). We find that a small but statistically significant portion of the interannual variance (>20 %) of the wintertime AO can be predicted at leads up to 2 months using lagged ensemble averages. As far as we are aware, this is the first study to demonstrate that an operational model has discernible skill in predicting AO variability on seasonal timescales. We find that the CFS forecast skill is slightly higher when a weighted ensemble is used that rewards forecast runs with the most accurate representations of October Eurasian snow cover extent (SCE), hinting that a stratospheric pathway linking October Eurasian SCE with the AO may be responsible for the model skill. However, further analysis reveals that the CFS is unable to capture many important aspects of this stratospheric mechanism. Model deficiencies identified include: (1) the CFS significantly underestimates the observed variance in October Eurasian SCE, (2) the CFS fails to translate surface pressure anomalies associated with SCE anomalies into vertically propagating waves, and (3) stratospheric AO patterns in the CFS fail to propagate downward through the tropopause to the surface. Thus, alternate boundary forcings are likely contributing to model skill. Improving model deficiencies identified in this study may lead to even more skillful predictions of wintertime AO variability in future versions of the CFS.

Keywords Arctic Oscillation · Stratosphere–troposphere coupling · Seasonal forecasting · Eurasian snow cover · Climate prediction · Modes of climate variability

This paper is a contribution to the Topical Collection on Climate Forecast System Version 2 (CFSv2). CFSv2 is a coupled global climate model and was implemented by National Centers for Environmental Prediction (NCEP) in seasonal forecasting operations in March 2011. This Topical Collection is coordinated by Jin Huang, Arun Kumar, Jim Kinter and Annarita Mariotti.

E. E. Riddle (✉) · A. H. Butler · A. Kumar
Climate Prediction Center, NCEP/NWS/NOAA, 5830 University Research Court, College Park, MD 20740, USA
e-mail: Emily.Riddle@noaa.gov

E. E. Riddle
Wyle Science Technology and Engineering Group, 1290 Hercules Ave., Houston, TX 77058, USA

J. C. Furtado · J. L. Cohen
Atmospheric and Environmental Research, 131 Hartwell Place, Lexington, MA 02421, USA

1 Introduction

The Arctic Oscillation (AO), characterized by opposing pressure anomalies in the Arctic and the northern midlatitudes, is the dominant mode of atmospheric climate variability in the Northern Hemisphere extratropics (Thompson and Wallace 1998). In the winter months, the positive (negative) AO is associated with warmer (colder) than average temperatures across northern Eurasia and the eastern United States, and colder (warmer) than average temperatures over northeastern Canada, Greenland and Alaska (e.g., Thompson and Wallace 2001; Buermann et al. 2003; Cohen and Barlow 2005). Given these widespread impacts on surface climate, skillful seasonal prediction of the December–February (DJF) AO index could improve wintertime climate outlooks such as those issued by the National Oceanic and Atmospheric Administration (NOAA) Climate Prediction Center (CPC).

Variability of the AO is considered to be primarily due to internal atmospheric dynamics and feedbacks (e.g., Robinson 2000; Lorenz and Hartmann 2003), making its prediction a challenge beyond a few weeks. However, some studies have suggested that a portion of the wintertime AO variability may be driven by external forcings that operate on longer timescales, permitting some predictability of the AO at leads of a month or more. For example, Cohen and Fletcher (2007) have examined the utility of using October Eurasian SCE as a predictor of the DJF AO. They find that 20 % of the variance in the DJF AO is explained by a simple SCE index describing the mean snow extent over Eurasia in October. This can be increased to 37 % of the variance when using a combination of the snow cover and sea-level pressure anomalies over Eurasia. A Snow Advance Index (SAI), which describes the rate that snow advances across Eurasia in October, can explain up to 74 % of the variance in the DJF AO, albeit for a short 14-year record (Cohen and Jones 2011a). Significant correlations have also been found between wintertime AO variability and a number of other predictors, including: (1) North Atlantic SSTs (e.g., Wang et al. 2004; Hu and Huang 2006), (2) the Quasi-Biennial Oscillation (QBO; e.g., Garfinkel and Hartmann 2007; Lu and Pandolfo 2011), (3) Solar variability (e.g., Ruzmaikin and Feynman 2002; Ineson et al. 2011), (4) late summer Arctic sea ice extent (e.g., Deser et al. 2010; Strong and Magnusdottir 2011; Liu et al. 2012), and (5) Pacific sea surface temperatures (SSTs; e.g., Garfinkel and Hartmann 2008; Ineson and Scaife 2009).

In order to understand these relationships, physical mechanisms are needed that operate on timescales of several months or more. Since the ocean tends to have a longer memory than the atmosphere, one possible mechanism involves the persistence of sea-surface temperatures anomalies from the boreal autumn into the winter season. For example, Derome et al. (2005) find that 20 % of the 1948–1998 DJF AO variance can be predicted with a dry atmospheric Global Circulation Model (GCM) forced with November oceanic anomalies persisted through the winter months. Using the same model, Tang et al. (2007) find that years with the largest oceanic anomalies in November tend to have the most predictable AO index in the following winter.

Mechanisms involving coupling between the stratosphere and the troposphere can also operate on monthly to seasonal timescales. AO-like geopotential height anomalies in the stratosphere, such as those associated with sudden stratospheric warming (SSW) events, can propagate downward into the troposphere and influence the surface AO on timescales of several weeks to 2 months (e.g., Baldwin and Dunkerton 1999, 2001). Furthermore, stratospheric anomalies (including SSWs) may be preceded by anomalous tropospheric precursor patterns (e.g., Garfinkel et al. 2010; Cohen and Jones 2011b), extending the timescale for prediction even further backward in time. A number of studies have proposed that these stratosphere-troposphere coupling mechanisms may explain observed relationships between various predictors (e.g., Pacific SSTs, solar variability, the QBO, and Eurasian SCE) and the wintertime AO (e.g., Cohen et al. 2007; Garfinkel et al. 2010; Fletcher and Kushner 2011). Accurate representation of stratosphere-troposphere coupling may improve dynamical model forecasts of Northern Hemisphere surface climate (Douville 2009; Orsolini et al. 2011; Sigmond et al. 2013).

The present study is divided into two parts. In the first part, we examine how well ensembles of the National Centers for Environmental Prediction (NCEP) operational Climate Forecast System model version 2 (CFSv2; Saha et al. in review) can be used to predict the DJF AO at lead-times ranging from 0 to 6 months. A few previous studies have examined the skill of CFSv2 at predicting seasonal climate variability (e.g., Yuan et al. 2011; Kim et al. 2012; Chen et al. 2013; Saha et al. in review), but none have focused specifically on seasonal predictions of the wintertime (DJF) AO index in CFSv2. We examine the sensitivity of our results to the ensemble size used, and to an ensemble member selection procedure that rewards runs that best track observed climate variables (e.g., Eurasian SCE) in the period before the forecast is made. This methodology is a simplified version of the "dynamic stratification" procedure defined by Schubert et al. (1992) for numerical weather prediction, and adapted for decadal climate prediction by Meehl et al. (2010).

In the second part of the study, we test how well the CFSv2 model represents troposphere-stratosphere-troposphere coupling mechanisms that might account for skill in the CFSv2 AO forecasts. We focus particularly on evaluating each step in a proposed mechanism connecting October Eurasian SCE to the DJF AO (e.g., Cohen et al. 2007), but the results are also applicable to other potential mechanisms involving interactions between the stratosphere and the troposphere. Our analysis is similar to that presented by Hardiman et al. (2008) and Furtado et al. (in revision) who evaluate the same mechanism in several coupled climate models as part of the Coupled Model Intercomparison project (CMIP), phases 3 and 5, respectively. However, by focusing on an operational model, we provide specific insights that will be useful for both users and for model developers working on the next generation of climate forecast models.

2 Methods

2.1 Data and models

This study uses seasonal retrospective forecasts (hindcasts) from the NCEP CFSv2 model, which consists of the NCEP

Global Forecast System (GFS) atmospheric model run at T126 (~100 km) horizontal resolution fully coupled with ocean, sea-ice, and land surface models (LSMs) (Saha et al. in review). The ocean model is the Geophysical Fluid Dynamics Laboratory (GFDL) Modular Ocean Model version 4.0 (MOM4; Griffies et al. 2004) at 0.25°–0.5° latitude by 0.5° longitude grid spacing. The LSM is the four-layer Noah LSM (Ek et al. 2003) and the sea ice model is an interactive three-level model. The CFSv2 forecast model is initialized using the atmospheric and surface fields from the Climate Forecast System Reanalysis (CFSR; Saha et al. 2010).

Four CFSv2 runs (00Z, 06Z, 12Z and 18Z) are initialized every 5 days starting on 12 December, 1981 and run for 9.5 months each. Monthly averaged output is used in this study. Since we focus on predictions of the DJF AO, we analyze only those runs that include the entire DJF timeframe in their output. For each winter season from 1982/1983 to 2009/2010, there are a total of 188 ensemble members initialized between 11 April and 27 November that cover the full DJF period.

The CFSR is used for the reanalysis in this study to calculate various atmospheric diagnostics and as a benchmark for the forecast simulations. The CFSR assimilates in situ and remote observations of the atmosphere, land-surface and oceans. It assimilates snow depth measurements from the Air Force Weather Agency's SNODEP model (Kopp and Kiess 1996) adjusted after 1997 to match snow cover data from the NOAA Ice and Snow Mapping System (IMS; Ramsay 1998).

The Northern Hemisphere snow cover observations used in this study are from the Rutgers Global Snow Lab (http://climate.rutgers.edu/snowcover/), provided on a 24-km equal area grid. From 1966 to 1997, these data are created based on the weekly satellite-based snow cover maps from the NOAA/National Environmental Satellite, Data, and Information Service (NESDIS) (Robinson et al. 1993). From 1997 to the present, the data are based on the daily NOAA IMS, which replaced the weekly NESDIS products (Ramsay 1998). The Rutgers record adjusts these NOAA satellite-based datasets to correct for discontinuities and biases in the 44-year record. For the Rutgers data, each 24-km grid cell is determined to be either entirely "covered" or entirely "snow-free" on any given day, and the monthly fraction is based on the number days in the month that the grid cell is covered. In contrast, the NOAH LSM used in CFSv2 allows each grid cell to be fractionally covered at each time step, with the snow cover fraction functionally related to the snowdepth. The monthly snow cover fraction in CFSv2 is the monthly average of the spatial fractions.

2.2 Indices and diagnostics

The AO index is calculated from an empirical orthogonal function (EOF) analysis of area-weighted monthly mean 1,000-hPa geopotential height anomalies poleward of 20°N, using the months of December–February 1982/1983–2009/2010. The monthly AO index (i.e., the time series associated with the leading mode of variability in the EOF analysis) is averaged to create a DJF index and then normalized with respect to its mean and standard deviation. The reference ("observed") DJF AO index is calculated using the CFSR geopotential height field and is very similar ($r = 0.995$) to the downloadable AO index from the CPC website (http://www.cpc.ncep.noaa.gov/products/precip/CWlink/daily_ao_index/ao.shtml). As on the CPC website, the AO loading pattern is defined as the regression of DJF 1,000 hPa height anomalies onto DJF AO index. In addition, 28-year DJF AO index time series are calculated for each of the 188 CFSv2 ensemble members by projecting the model 1,000 hPa geopotential height anomalies onto the CFSR AO loading pattern and normalizing with respect to their mean and standard deviation. CFSv2 anomaly fields are calculated with respect to their lead-dependent climatologies.

We calculate an October Eurasian SCE index using a method intended to achieve as much consistency as possible between the model and observations, despite different grids and land masks. Both the model and observed October snow cover records are first converted to a T62 Gaussian grid and adjusted manually to account for land mask differences. Eurasian SCE is then calculated as a weighted sum of the T62 grid cell areas, with weights determined by the monthly snow cover fraction at each grid cell. Northern Hemisphere grid cells between 26°W and 190°E are used in the index calculation, with Greenland excluded. The resulting observed Eurasian snow cover extent (SCE) data matches very closely to the monthly SCE data provided by the Rutgers Snow Lab.

2.3 Ensemble averaging

As mentioned previously, 188 runs in the CFSv2 hindcasts dataset provide simulations of wintertime climate each year. In general, the most recent of these runs (those initialized in November) should have the most skill, while the oldest runs (those initialized during the previous spring) would typically be discounted by forecasters, since the model forecasts eventually lose information about the initial atmospheric and oceanic conditions and converge towards the model climatology. The older runs may be useful for some purposes, however. First, from a forecasting perspective, adding ensemble members generally improves model skill by eliminating the contribution from the noise component (e.g., Déqué 1997; Kumar and Hoerling 2000; Kharin et al. 2001; Kumar et al. 2001; Wilks 2011; Chen et al. 2013). While including the oldest

runs will likely degrade rather than improve the prediction skill, it is not obvious at what point the advantage of a larger ensemble is outweighed by the disadvantage of adding less skillful runs (Chen et al. 2013). Second, from a model diagnostics perspective, the large ensemble can be used to examine mechanistic relationships internal to the model. In the second case, the quality of the model initialization is less important than having a large number of internally consistent model runs to evaluate.

In the first part of the paper, we test different methods of averaging the available ensemble members to make predictions of the DJF AO index at leads of 0–6 months. The forecast lead refers to the number of months (or days) between the date the forecast is made and the first day of the forecast period, in this case 01 December. Table 1 lists the number of ensemble members available for various leads. Instead of making any prior assumptions about the optimal number of ensemble members to use, we examine the skill of AO forecasts at different leads based on different-sized ensemble averages that selectively retain some or all of the older runs. For each of these ensemble averages, the predicted DJF AO index is calculated by averaging the AO indices (calculated as the projection onto the CFSR AO loading pattern) from all the desired ensemble members. The resulting 28-year ensemble-mean forecast time series is then detrended and normalized with respect to its mean and standard deviation. The forecast skill is evaluated based on the anomaly correlation between the predicted and the observed (CFSR) DJF AO time series for 1982/1983–2009/2010.

The second part of this study uses the available ensemble members to diagnose physical mechanisms internal to the CFS model. Here, we analyze runs that capture the full period covering October through February to see how the CFS translates model anomalies in October and November into wintertime climate variations. Each year, there are 140 runs available for this analysis, initialized between 11 April and 28 September. Each lead-time is considered as a separate "run", and diagnostics are performed on each run separately. Diagnostics, including lagged correlations, are then averaged over the 140 runs. We note the fundamental difference between this approach and the methodology in the first half of the study in which the ensemble averaging occurs before the correlation is calculated.

3 Prediction of the DJF AO

In this section, we evaluate CFSv2 forecasts of the DJF AO as a function of the forecast lead and ensemble size. For this analysis, an ensemble average of N members always uses the most recent N runs available on the date that the forecast is issued. All of the figures in this section show anomaly correlations between predicted and observed AO time series, as described in Sect. 2.3.

Figure 1a examines the skill of 4-, 20- and 64-member ensemble averages, as function of lead-time. The horizontal grey line indicates the threshold correlation value associated with statistical significance at the 5 % level, as determined by a two-tailed Student t test. As would be expected, the correlation values are generally larger for short-lead forecasts compared with long-lead forecasts. The 4-member ensemble averages show considerable variability in skill from one lead to the next, due to random variations in the correlation values as a result of the relatively short 28-year time series (Kumar 2009). Individual peaks (e.g., at 74, 99 and 154 days) are the expected result of this noise, and are not likely to be meaningful increases in skill. Only a handful of the 4-member forecasts pass a significance test at the 5 % level, however even the very long-range forecasts (4–6 month lead) show more positive correlations than negative correlations which possibly indicates some skill, even at these long leads. The 20- and 64-member ensemble averages are smoother functions of the lead-time, with the 64-member ensemble showing a monotonic increase in skill as the lead-time decreases. Correlation values exceed the 5 % significance threshold at leads of 3.5 months or less in the 64-member averages.

Figure 1b displays the effect of ensemble size on the 0, 1, 2, 3 and 4-month lead forecasts. Note that the larger ensemble sizes include older, less-skillful forecast runs in addition to the newer, more skillful ones. All ensemble sizes of 12 members or more produce significant correlations for the 0-, 1- and 2-month lead forecasts. Some ensemble sizes also produce significant correlations for the 3- and 4-month forecasts. The model skill improves initially with the inclusion of older model runs, but then flattens or decreases slightly as much older runs are added.

Table 1 CFSv2 runs available each year for the prediction of the DJF AO

	0-month lead	1-month lead	2-month lead	3-month lead	4-month lead	5-month lead	6-month lead
Date of forecast	Nov 30	Oct 31	Sept 30	Aug 31	July 31	June 30	May 31
Number of runs available	188	164	140	116	92	68	44
Range of model initialization dates	11 Apr–27 Nov	11 Apr–28 Oct	11 Apr–28 Sep	11 Apr–29 Aug	11 Apr–30 Jul	11 Apr–30 Jun	11 Apr–31 May

Figure 1c, d are similar to Fig. 1b except showing results for the first and second halves of the record separately. While positive correlations exist for both periods, the skill is mainly seen in the most recent 14 years (1997–2010) where correlations as high as 0.8 are seen in the 0-month lead forecasts, and correlations between 0.6 and 0.7 are observed for the 1- and 2-month lead forecasts. This difference between the first and second halves of the record may be related to systematic errors and biases prior to 1998 that have been noted in the initialization of the CFSv2 hindcasts (Saha et al. in review). A large positive model bias in October Eurasian SCE also exists prior to 1997, as will be discussed in Sect. 4.

One intriguing aspect of Fig. 1b–d is that very large ensembles (i.e. 40–70 members), including older runs initialized as early as July, appear to have skill which is comparable to or better than smaller ensembles using only the most realistic recent runs. Chen et al. (2013) recently investigated the trade-offs in a lagged ensemble between the benefits of increasing the ensemble size and the drawbacks of including longer lead forecasts. Like in Fig. 1b–d, they found that the model skill improves initially with the inclusion of older model runs, but eventually drops off, so that an "optimal lagged ensemble time" can be calculated where the skill is maximized. They found that the optimal ensemble size is highly dependent on the geographic location and forecast variable, with the optimal number of ensemble members ranging from approximately 8 members for tropical SSTs to approximately 60 members for extratropical precipitation. Variables with low predictability generally benefitted from using the largest number of lagged ensemble members, consistent with the theoretical results of Kumar and Hoerling (2000).

Given sampling errors in the forecast skill (Kumar 2009), and the fact that we are only forecasting a single index value, and therefore cannot average over multiple grid cells or seasons, it is not straight forward to determine the optimal lagged ensemble for our wintertime AO forecasts. Doing this would require fitting a statistical model to estimate underlying parameters of the model/climate system (e.g., underlying model skill, predictability) as a function of the lead-time and then using these parameters to determine a theoretical optimal ensemble size. Even without these steps, however, we can say qualitatively that a very large 64-member lagged ensemble appears more skillful at many lead-times than a smaller 20-member ensemble, despite including many older runs (Fig. 1a). As such, forecast skill of the wintertime AO, especially at leads of 1–3 months, might be improved by expanding the ensemble beyond the 10–20 members that are typically used.

Figure 2a is a generalization of the plots in Fig. 1, examining both the effect of different forecast leads and different ensemble sizes. The *x*-axis gives the initialization date of the earliest run used in the ensemble average, while the *y*-axis gives the initialization date of the latest run. All

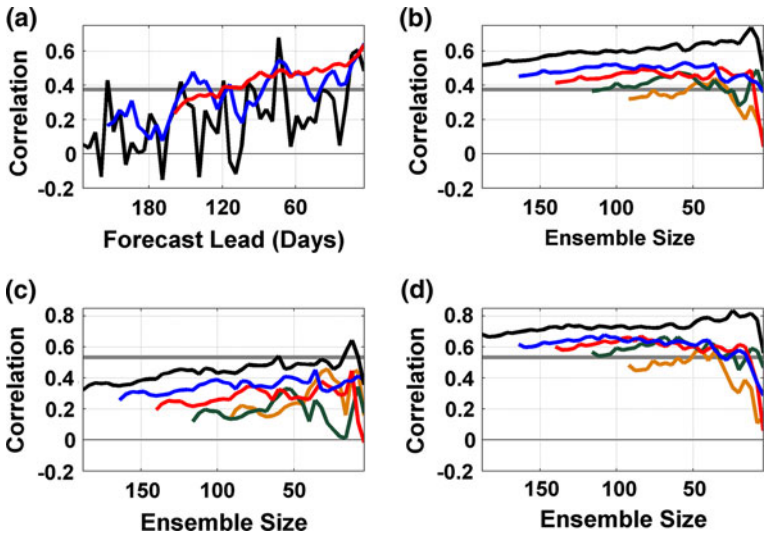

Fig. 1 **a** Correlations between the predicted DJF AO index and the observed DJF AO index time series at different forecast leads using the most recent 4 runs (*black*), 20 runs (*blue*), and 64 runs (*red*) in the ensemble average. **b** Skill associated with the 0-month lead (*black*), 1-month lead (*blue*), 2-month lead (*red*), 3-month lead (*green*) and 4-month lead (*gold*) forecasts as a function of the ensemble size. **c** Same as **b** except using only the first 14 years of the record before 1997. **d** Same as **b** and **c** except using only the second half of the record, from 1997 onward. The small ensembles only use the most recent runs, while the larger ensembles include older runs as well. *Horizontal grey lines* show the threshold for statistical significance at the 5 % level. All time series are detrended before the correlations are made

ensemble members in between these start and end dates are included in the forecast. As such, the diagonal of Fig. 2a is equivalent to the correlation values plotted in Fig. 1a (black line), and the top-most row is equivalent to the correlation values in Fig. 1b (black line). What is most striking about Fig. 2a is that positive correlations are observed for the large majority of these different ensemble averages, even those including only much older runs.

As shown in Table 1, there are 164 ensemble runs available to use in a 1-month lead forecast (i.e. a forecast made at the end of October). In addition to the model runs, the forecaster has access to the actual climate conditions that occurred in the months previous to the forecast, including the state of ENSO and the autumn Eurasian snow cover progression. This information could theoretically be used to discard earlier runs that are the least successful at capturing the observed climate conditions, and retain earlier runs that are the most successful at representing the observed conditions. This approach is similar to the "dynamic stratification" procedure defined by Schubert et al. (1992) for numerical weather prediction, and adapted for decadal climate prediction by Meehl et al. (2010).

Figure 2b–d apply dynamic stratification to the CFSv2 hindcasts based on October Eurasian SCE. The month of October is chosen because previous studies have found October SCE to be an important predictor of the wintertime AO (e.g., Cohen and Fletcher 2007; Cohen et al. 2007), and because October SCE data are available by the time of the 1-month lead forecast for wintertime (DJF) climate, a useful lead-time for seasonal forecasting. Figure 2b is similar to Fig. 2a in that the dates on the x and y axes define a window of initialization times to be included in the ensemble mean. However, instead of averaging all of the runs in that window, only those that best capture the actual observed October Eurasian SCE are retained, as determined by the smallest absolute difference between the observed October Eurasian SCE and the modeled SCE. More specifically, half of the runs in the specified window initialized before 1 October are retained and half are discarded based on their absolute SCE error. All runs initialized after 1 October are retained since October SCE data are not available for these runs. As we will see in Sect. 4, the model tends to have a high bias in October Eurasian SCE, especially in the earlier years of the record (before

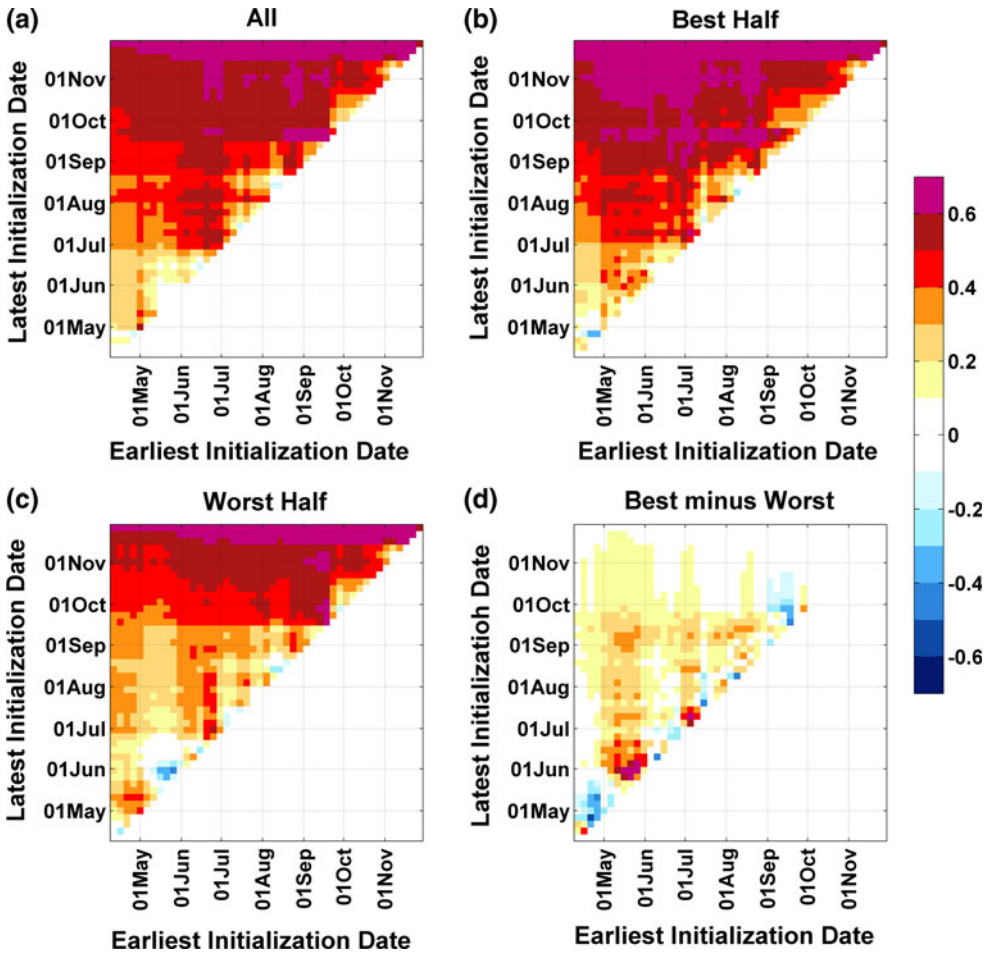

Fig. 2 **a** Correlations between the predicted DJF AO index and the observed DJF AO index time series using ensembles of different lengths. The earliest and latest initialization dates are bounds on the dates used in the ensemble average. **b** As in **a** except only ensemble members with the best October Eurasian SCE each year are used in the average. **c** As in **a** and **b** except only using ensemble members with the worst October Eurasian SCE. **d** Difference between the correlation values in **b** and **c**

1997). Thus, the best runs in early years are often those with the lowest absolute snow extents, while after 1997 an average of the best runs tends to match the highs and lows of the observed snowfall variations quite well.

Figure 2c is similar to Fig. 2a but shows the skill when retaining runs with the "worst" October Eurasian SCE. Figure 2d illustrates the difference between Fig. 2b and Fig. 2c. In most cases, we see an increase (decrease) in skill for the forecasts that use only the best (worst) October Eurasian SCE values (Fig. 2b, c), in agreement with the October Eurasian SCE–wintertime AO hypothesis. If we plot the results from Fig. 2d separately for the first and second halves of the record (not shown), we find that the improvement in AO prediction occurs only for the second half of the record (1997–2010) when the "best" runs more accurately represent October Eurasian SCE variability. We note that the specific choices used in this analysis (e.g., retaining 50 % of the applicable runs, and stratifying based on the absolute error in October Eurasian SCE) have not been optimized, and further work is needed to examine whether other choices, including retaining more or fewer runs, would improve the results.

Figure 3 extends the analysis presented in Fig. 2 to test stratification based on other climate variables. The top-most panels of Fig. 3 again show the results of dynamic stratification based on October Eurasian SCE, as in Fig. 2. The left-hand panels show the 1-month lead forecast skill for ensemble forecasts using all members, the best members and the worst members (as in Fig. 2a–c, respectively), while the right-hand panels show the difference in skill between the "best" and "worst" ensembles (as in Fig. 2d). Ensemble members are selected from a window of initialization dates ranging from the date on x-axis label to 28 October when the latest runs available for the 1-month forecast are initialized. As in Fig. 2, the 24 ensemble members initialized during October are included in all ensembles and not subject to stratification. The statistical significance of the results is determined based on a resampling test where the "best" and "worst" 50 % of ensemble members are chosen at random from the available pool of runs. Percentiles of this null distribution are calculated from 10,000 random selections, and are plotted in grey on the right-hand panels. There is generally improvement associated with using only the best runs based Eurasian SCE. The improvement exceeds the 90th percentile of the null distribution for some ensemble averages, but always falls below the 95th percentile and so cannot be considered statistically significant at the 5 % level.

Figure 3c–f are similar to Fig. 3a, b, but use October SSTs in the Nino 3.4 region (Fig. 3c, d) and the October AO index (Fig. 3e, f) instead of October Eurasian SCE as criteria to select the "best" runs. These selections perform no better than October Eurasian SCE at distinguishing the best from the worst ensemble members and also fail to exceed the 5 % significance threshold. In fact, of the three variables tested, October Eurasian SCE is the only one where stratification based on the best members shows considerable improvement over using all members for the majority of ensemble sizes. The results from Figs. 2 and 3 are inconclusive, but suggest that runs with realistic representation of October Eurasian SCE may lead to better CFSv2 ensemble forecasts of the DJF AO.

4 CFSv2 model diagnostics

In the previous section, we demonstrated that CFSv2 ensemble forecasts of the DJF AO have modest but positive skill at leads up to a few months, raising questions about potential sources of skill in the CFS model. Furthermore, dynamic stratification based on October Eurasian SCE hinted that an accurate representation of Eurasian SCE might be important to model skill. While a full diagnosis of the sources of the model skill is beyond the scope of this paper, this section tests one potential mechanism: a stratospheric pathway linking October Eurasian SCE variability to the DJF AO (Cohen et al. 2007). This theoretical mechanism begins with a tropospheric response in October and November to anomalous snowfall covering the landscape over Eurasia. A characteristic upstream ridge over the north Pacific and downstream trough over Eastern Europe develop in response to the SCE anomaly. The location of these anomalies serves to amplify the climatological wave 1 and 2 patterns in the troposphere, leading to increased vertical planetary wave propagation into the stratosphere in November and December. These waves break in the stratosphere, slowing the polar vortex, warming polar stratospheric temperatures and leading to a negative stratospheric AO signal during the early winter. The stratospheric anomalies then propagate down into the troposphere, resulting in a negative surface AO tendency in DJF.

We will attempt to diagnose the CFS model ability to capture each of the steps in this pathway using the 140 CFSv2 ensemble members that capture the full October–February period (i.e., the period covering the full troposphere-stratosphere-troposphere pathway linking October SCE anomalies to the wintertime AO). The results will be applicable to other potential AO-forcing mechanisms besides October Eurasian SCE, since other proposed mechanisms also involve troposphere–stratosphere interactions. For example, Pacific SST anomalies associated with ENSO can also create a persistent trough over the north Pacific (e.g., Garfinkel et al. 2010, 2012; Fletcher and Kushner 2011; Hurwitz et al. 2012), initiating a similar response in the troposphere and stratosphere as Eurasian

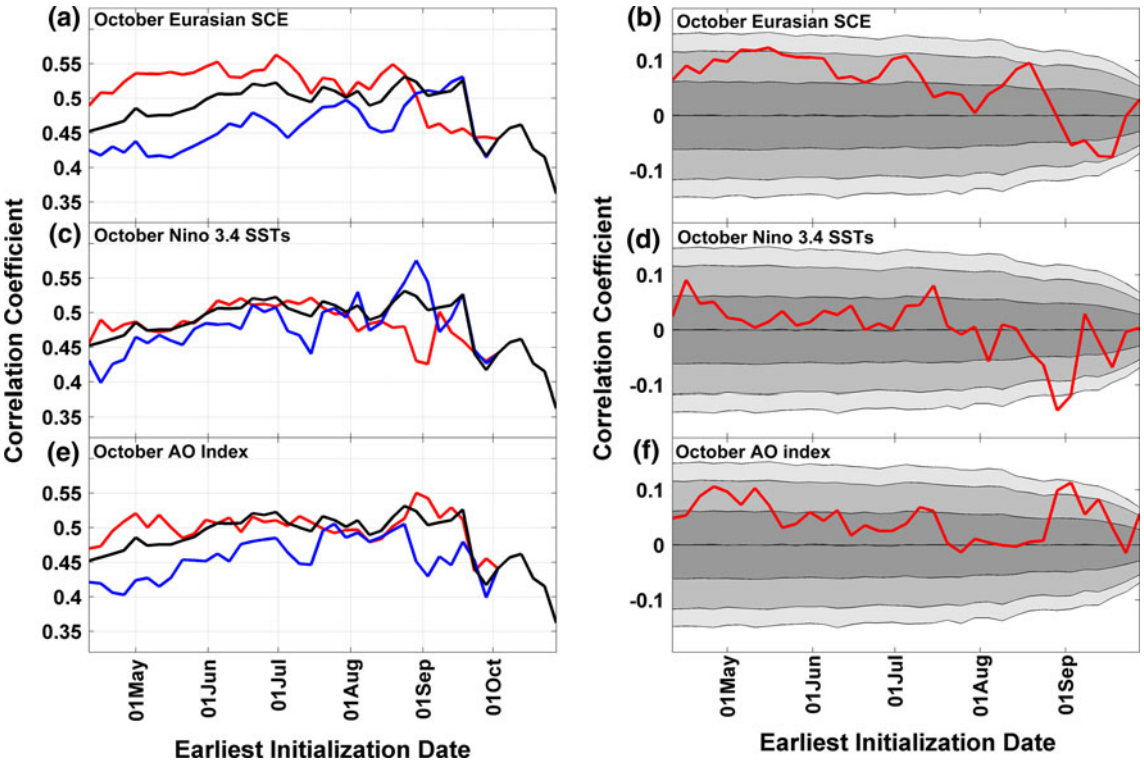

Fig. 3 **a** The 1-month lead forecast skill in predicting the DJF AO using all ensemble members (*black line*), using only those members with the best representation of October Eurasian SCE (*red line*), and using only members with the worst October Eurasian SCE (*blue line*; see text for details). **b** (*Red line*) The difference in skill between the best and worst October Eurasian SCE ensemble members (i.e., the *red and blue lines* in **a**). *Thin black lines* with *grey shading* show from *bottom* to *top*, the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles for a null distribution where the best half of runs are chosen at random. **c** As in **a**, except the best half and worst half are chosen based on ensemble members with the best October SSTs in the Nino 3.4 region. **d** As in **b**, except showing the difference between the *red* and *blue lines* in **c**. **e** As in **a**, except that the best half of ensemble members is chosen based on the best October AO index values. **f** As in **b**, except showing the difference between the *red* and *blue lines* in **e**

SCE anomalies. The following four sub-sections will focus on different links the chain of events described above. The first sub-section will examine how well the model runs reproduce the DJF AO loading pattern. The second sub-section will examine how well the model runs reproduce variability in Eurasian SCE. The next sub-section will look at the surface response to snow cover anomalies. The final sub-section will diagnose the remaining stages of the troposphere-stratosphere coupling mechanism.

4.1 The Arctic Oscillation

Figure 4 examines how well the CFSv2 model runs reproduce the wintertime AO spatial loading pattern. The CFSR AO loading pattern is shown in Fig. 4a. It resembles the AO pattern identified in many previous studies (e.g., Thompson and Wallace 1998) except that the Atlantic and Pacific sectors show approximately equal loadings, whereas other studies have found a stronger loading in the Atlantic. Figure 4b shows the mean CFSv2 loading pattern, averaged over all 140 runs initialized between April and September. We omit runs initialized after October 1 (i.e., those with leads less than 2 months) in order to focus on the same set of runs that will be used in the remainder of this section to study the atmospheric response in CFSv2 to modeled October SCE anomalies. To calculate the CFSv2 AO loading pattern, first a run-specific AO index is computed using an EOF analysis of December–February 1,000 hPa geopotential height anomalies. Then, the run-specific loading pattern is calculated as the regression of monthly 1,000 hPa height anomalies onto this index. If the AO is not the leading mode in the EOF analysis, the next two modes are also tested and the mode that most resembles the observed AO pattern is used. The AO is the leading pattern in 121 out of the 140 runs. We have also calculated the AO loading pattern from an aggregated sample of all months (December–February), runs and years with very similar results (not shown).

Qualitatively, the mean CFSv2 loading pattern (Fig. 4b) is very similar to the CFSR loading pattern (Fig. 4a) , though the regression coefficients are slightly weaker on average and the negative pole lacks an

extension over northern Eurasia. The weaker signal may be due to averaging over the 140 runs. The loading pattern calculated based on the aggregated sample has slightly stronger magnitudes, especially over the Pacific sector (not shown). Figure 4c shows the distribution of pattern correlations between the loading patterns from each of the 140 runs and the observed loading pattern. Most are relatively well correlated, though the pattern correlation is below 0.5 for a few runs, suggesting that a canonical AO-like pattern does not always emerge. Because of this, the CFSv2 AO index used in the remainder of the paper is the projection of CFSv2 height anomalies onto the CFSR loading pattern, ensuring that the index represents the canonical AO.

4.2 October Eurasian snow cover

Figure 5a, b show the mean 1982–2009 climatology of October Eurasian snow cover percent based on the satellite-based Rutgers observational dataset (Fig. 5a), and based on an ensemble average of the 140 CFSv2 model runs (Fig. 5b). The difference between Fig. 5b and a is shown in Fig. 5c. The CFSv2 model runs tends to produce too much snow cover in October over a band centered around 60°N, suggesting that central Russia is likely covered earlier in October in the model compared with the observations. Too little snow is simulated over some regions of mountainous northeastern Siberia, though discrepancies in this region may be partly due to higher resolution in the original model grid compared with the Rutgers dataset. Figure 5d shows October Eurasian SCE time series from the observations and the CFSv2 140-member ensemble average. Note that the ensemble average shows little year-to-year variability and a slight downward trend, while the observations show much larger variations and an upward trend. Due to the trend in the observations, the CFSv2 SCE bias is much higher in the second half the record compared with the first half of the record.

Figure 6 shows the ensemble distribution of the mean and standard deviation of October Eurasian SCE in the 140 CFSv2 runs compared with the observations. In general, the climatological snow extent is too large in the model runs (though this bias is mostly limited to the first half of the record prior to 1997 as shown in Fig. 5d), while the standard deviation is much too small. Figure 6a, b indicate that none of the 140 model runs have a 28-year mean October Eurasian SCE that is as small as the observed mean SCE, or a year-to-year variance that is as large as the observed SCE variance. In some of the runs, the standard deviation is less than half the observed standard deviation. Figure 6c, d show the same data as Fig. 6a, b, but with the model runs plotted as a function of the lead before October 1. The runs with the shortest leads show the most realistic

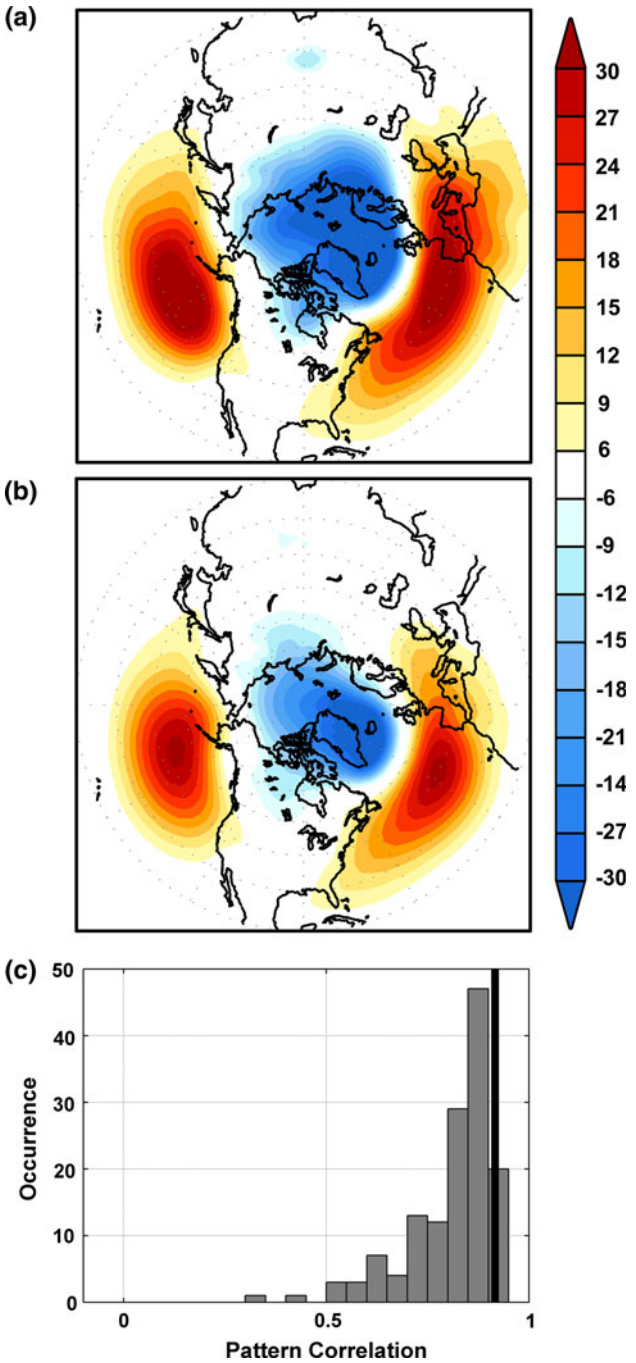

Fig. 4 **a** Regression map of 1,000 hPa October–February geopotential height anomalies (m) onto the CFSR AO index. **b** Same as **a** except that the plot is an average of the 140 regression maps obtained separately for each of the 140 CFSv2 runs. **c** Histogram showing the distribution of pattern correlation values between the CFSR regression map in **a** and regression maps obtained individually for each of the 140 CFSv2 hindcast runs. The *vertical black line* shows the pattern correlation value between the regression maps shown in **a** and **b**

(i.e. lowest) mean SCE and also tend to have slightly more realistic (i.e. higher) standard deviations. Intermediate lead-times of approximately 1–2 months, however, show

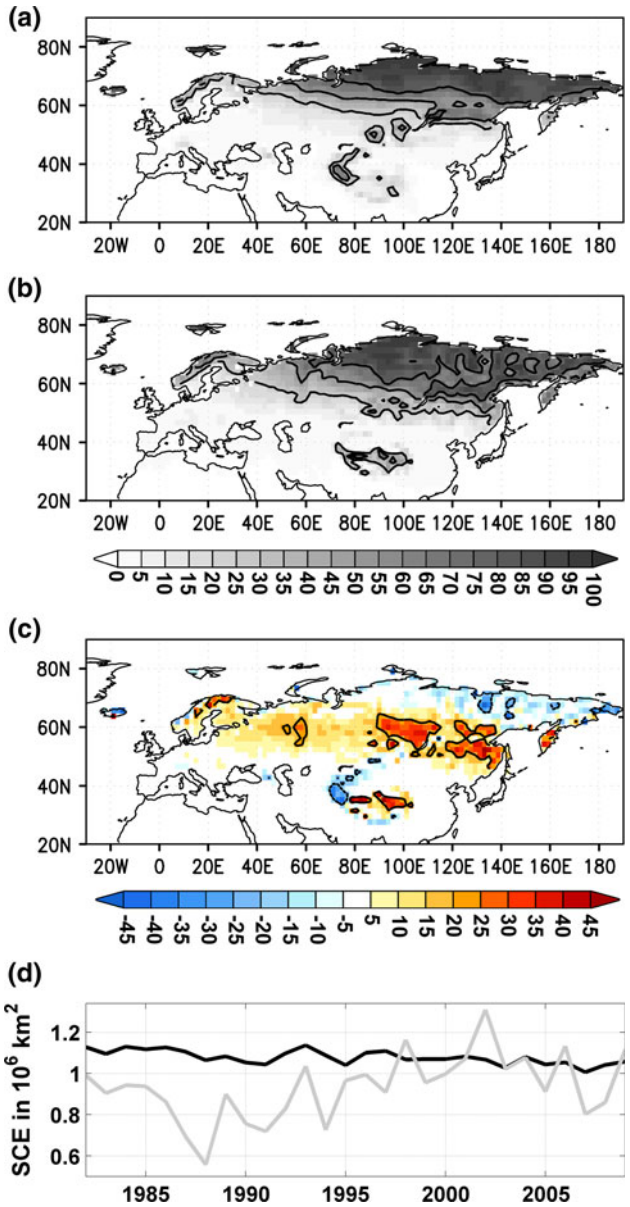

Fig. 5 Percentage of days in October from 1982 to 2009 that a grid cell is covered with snow **a** in the Rutgers observations and **b** in the 140-run ensemble average of the CFSv2 hindcasts. *Solid lines* show the 25, 50 and 75 % contours. **c** Difference in snow cover percentage between the CFSv2 model and the Rutgers observations. *Solid lines* show differences of +25 and −25 %. **d** Time series of Rutgers October Eurasian SCE observations (*grey*) and CFSv2 140-member ensemble average October Eurasian SCE (*black*)

the least realistic representations of the mean snow extent. This may be related to spin-up problems in the model at these intermediate lead-times.

The CFSv2 SCE statistics presented here are consistent with previous studies, which have found that models tend to underestimate the year-to-year variability of snowfall (e.g., Hardiman et al. 2008; Allen and Zender 2010; Furtado et al. in revision). The lack of snowfall variability has been implicated as one of the primary reasons that models with simulated snow cover fail to capture relationships between October snow cover and the wintertime AO. In contrast, models that have been forced with either idealized or observed snow cover tend to better simulate the proposed mechanism (Cohen and Entekhabi 2001; Gong et al. 2003; Fletcher et al. 2009; Orsolini and Kvamstø 2009; Allen and Zender 2010, 2011; Peings et al. 2012), suggesting that better representation of snow cover variability may be important for capturing the SCE/AO relationship.

Recent work has suggested that the latitude at which the snow cover forcing occurs may be important. Cohen and Jones (2011a) find that the strongest relationship between October Eurasian SCE and the DJF AO occurs when considering SCE variability late in October and south of 60°N. Because Eurasian SCE is consistently too high in the CFSv2 model, key forcing regions may always be covered by the end of October, resulting in even less year-to-year variability in these areas.

4.3 The surface response to October Eurasian SCE

The response of the surface geopotential height field to October Eurasian snow cover is shown in Fig. 7. The first column shows the difference in CFSR 1,000 hPa geopotential height anomalies for the 7 years (i.e. 25 % of years) with the highest observed October Eurasian SCE minus the 7 years with the lowest observed October Eurasian SCE. The second and third columns show composite anomaly maps associated with high-snow minus low-snow CFSv2 runs, with the high-snow and low-snow runs drawn from the aggregated pool of all CFSv2 ensemble members and years (i.e. a pool of $28 \times 140 = 3{,}920$ runs). The second column shows seven-sample composite anomalies, matching the observed sample size, while the third column shows 980-sample composite anomalies, created from the top and bottom 25 % of runs in the aggregated pool. Histograms in the last column show a distribution of pattern correlations between the observed composite anomalies and model composite anomalies using seven high-snow and seven low-snow runs, drawn at random from the top/bottom 25 % of the aggregated pool of runs.

In October, positive SCE anomalies are associated with a negative AO/NAO pattern, with positive geopotential height anomalies over Greenland, the North Pole and eastern Siberia, and negative anomalies over the North Atlantic, northern Europe and western and central Russia. The model shows a similar signal to the reanalysis in October, as seen in both the small-sample (Fig. 7b) and the large-sample (Fig. 7c) composite differences. Note that the SCE difference between the high-snow and low-snow composites is slightly larger in the 7-sample CFSv2 composites than in the observations since only the runs with the

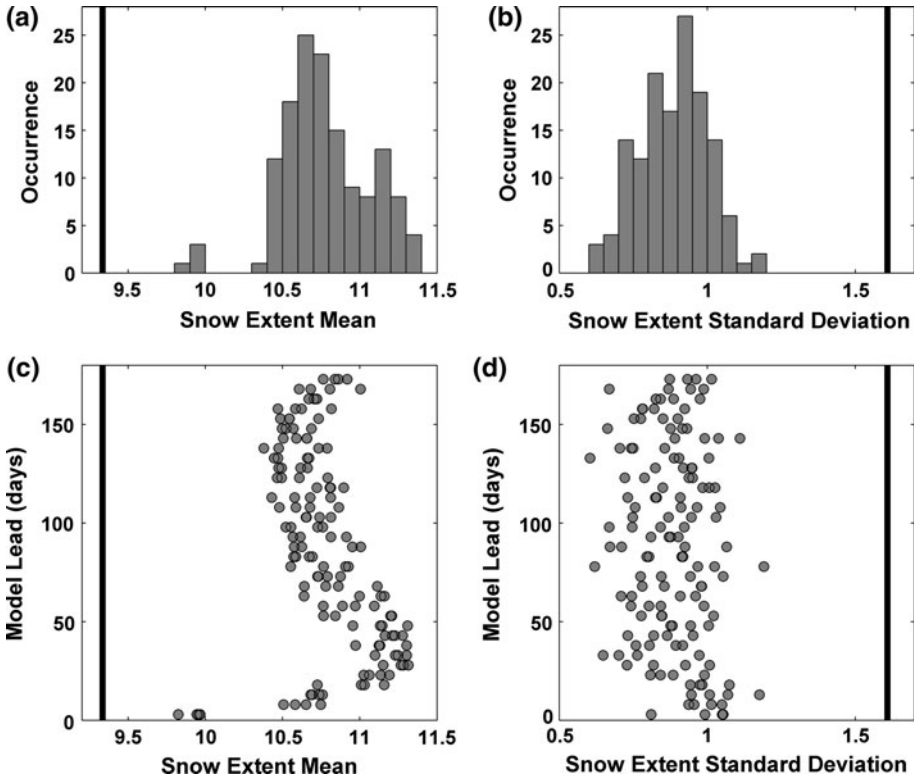

Fig. 6 *Top* Histograms showing ensemble distributions of the **a** climatological mean and **b** standard deviation of October Eurasian SCE for 1982–2009 both in units if millions of square kilometers. *Bottom* Same as *top* except showing the dependence of the **c** climatological mean and **d** standard deviation on the model lead in days before 1 October. The *vertical black lines* show the 28-year mean (**a**, **c**) and standard deviation (**b**, **d**) of observed October Eurasian SCE

very highest and very lowest SCE are selected, but almost a factor of two smaller in the 980-sample composites than in the observations (see Fig. 7 caption). While variability due to sampling is quite large, pattern correlations in Fig. 7d also suggest that 7-sample CFSv2 anomalies are mostly consistent with the observed anomaly pattern in October, with positive pattern correlations between the two occurring in more than 93 % of the 10,000 random draws. Because the snow anomalies are contemporaneous with the geopotential height signal in October, it is not possible to determine if the geopotential height signal causes the snowfall anomalies or vice versa.

During November and December, the positive anomalies in CFSR geopotential height extend southward and westward across much of northern Eurasia, and a trough develops over the northeastern Pacific (Fig. 7e, i). This Pacific trough has been identified as an important feature driving increased wave activity fluxes into the stratosphere (e.g., Garfinkel et al. 2010; Hurwitz et al. 2012). In contrast, the 7-sample CFSv2 composite anomalies for November and December show a very different pattern, including a strong ridge over the northeastern Pacific (Fig. 7f, j), likely due to sampling variations. The large-sample composite anomalies lack any strong coherent signal in November and December, though a very weak trough may be detected over the northeastern Pacific (Fig. 7g, k), consistent with the sign of the observed signal but more than an order of magnitude weaker. The pattern correlation histograms (Fig. 7h, l) also demonstrate only a very weak tendency towards positive correlations.

In January and February, the CFSR geopotential height anomalies become larger, more annular, and begin to resemble a canonical negative AO pattern over the Atlantic, North American and Eurasian sectors. The 7-sample CFSv2 anomalies show little resemblance to the observed anomalies with any patterns likely due to sampling variability. No signal is present in the large-sample model anomalies or the pattern correlation histograms. These results indicate that the CFS model, similar to other models studied (e.g., Hardiman et al. 2008; Furtado et al. in revision), loses the surface response to snow cover anomalies in the months following October.

We have also examined surface temperature anomalies associated with SCE variability (not shown). In October, high Eurasian snowfall is associated with strong negative temperature anomalies over most of Siberia with the highest anomalies over the eastern portion of the continent. Weaker but still significant positive temperature anomalies are observed over Greenland and central Asia. The model runs show a similar surface temperature signal, but with location of the negative anomalies over western instead of eastern Siberia.

Figure 8 shows correlations between the October Eurasian SCE signal and the DJF AO. Correlation values between these two detrended time series are calculated separately for each of the 140 runs and shown as a

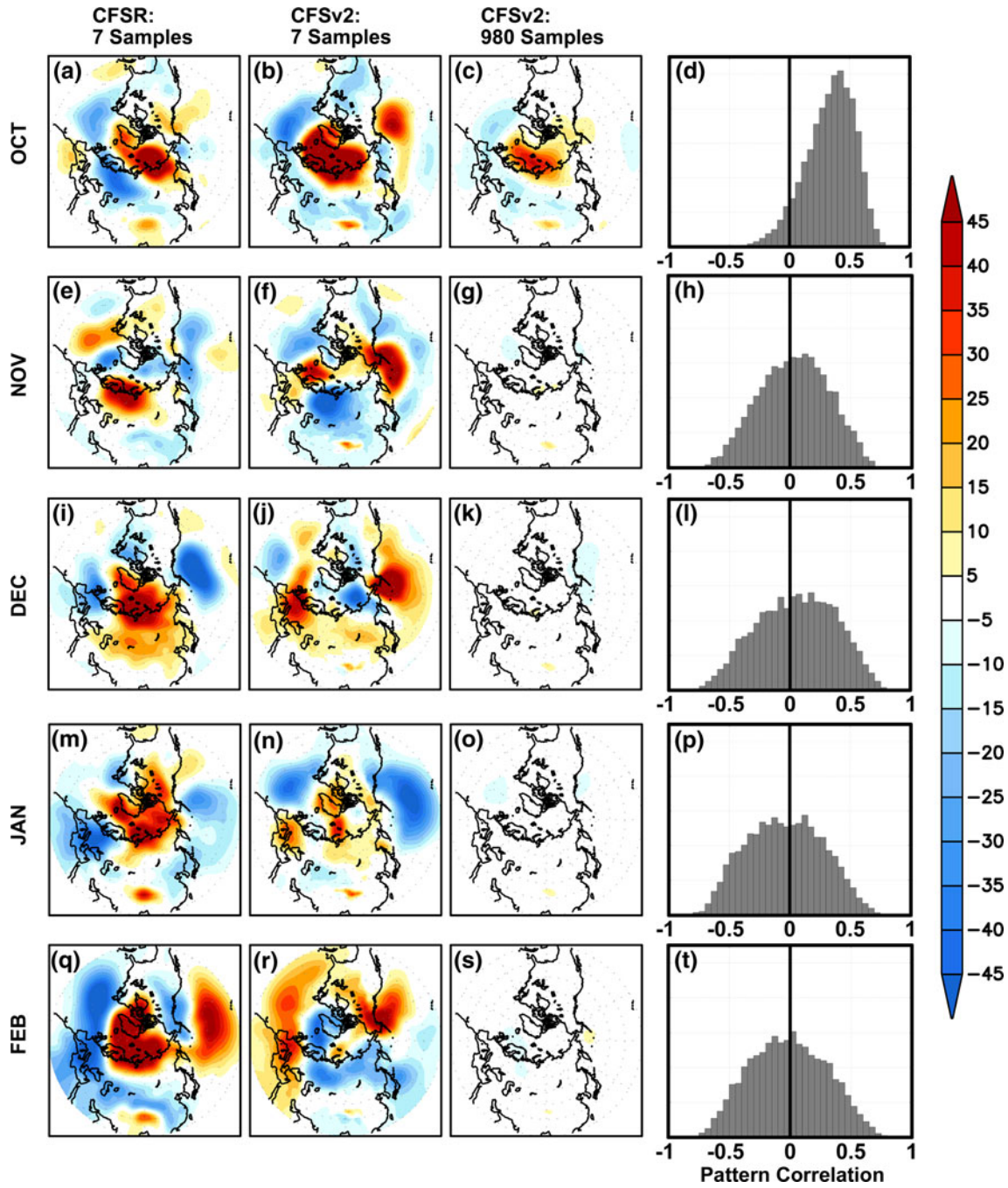

Fig. 7 **a** Composites of October 1,000 hPa CFSR geopotential height anomalies (m) for the 7 years with the highest observed October Eurasian SCE (*top* 25 % of cases) minus the 7 years with the lowest observed October SCE (*bottom* 25 % of cases). Average October SCE is 11.3 (7.3) million square kilometers for these high (low) years. **b** Composites of October 1,000 hPa CFSv2 height anomalies for the 7 cases in the CFSv2 model runs with highest October Eurasian SCE, minus the 7 cases with the lowest modeled October Eurasian SCE. Average October SCE is 13.9 (7.9) million square kilometers for these high (low) cases. Cases are drawn from the combined pool of all ensemble members and years. **c** Composites of October 1,000 hPa CFSv2 height anomalies for the 980 cases in the CFSv2 model runs with highest October Eurasian SCE (*top* 25 % of cases) minus the 980 cases with the lowest modeled October Eurasian SCE (*bottom* 25 % of cases). Average October SCE is 11.9 (9.6) million square kilometers for these high CFS (low) cases. **d** Histogram of anomaly correlations between *panel* **a** and 10,000 composite maps calculated from the difference between 7 high-snow and 7 low-snow cases from the CFSv2 hindcasts. The high snow and low snow cases are chosen at random from pools of the *top* and *bottom* 25 % of cases. **e–t** Same as *above panels* except for geopotential height anomalies for November–February

histogram in Fig. 8a and as a function of lead-time in Fig. 8b. The observed correlation value is indicated with vertical black lines and the null distribution for correlations between two unrelated 28-year time series is shown in red. The histogram is centered near zero with a mean correlation of −0.002 and a standard deviation of 0.188, and is not statistically distinguishable from the null distribution, which has a mean of zero and a standard deviation of 0.193. Correlations do not tend to improve at shorter leads. We have repeated Fig. 8 using only the most recent 14 years of the record (not shown). The mean correlation is −0.054 in the recent record, which is statistically significant at the 5 % level due to the large number of degrees of freedom (140 ensemble members and 14 years), but still very weak.

More work is needed to reconcile these results with those in the previous section. If snow cover in the model is so weakly correlated with the DJF AO, why do runs with the best representations of October Eurasian SCE create better forecasts of the DJF AO? One possibility is that runs with unrealistic October Eurasian SCE have other deficiencies that make them unsuitable for wintertime AO prediction. On average, the ensemble members with the most realistic October Eurasia SCE do have slightly higher covariance between snow anomalies and negative AO anomalies, though the difference is small and not statistically significant.

Given the relatively short length of the observational record and deficiencies in the model, it is not possible at this time to conclusively determine the "true" correlation, ρ (e.g., the correlation if we had an infinite, stationary observational record) between the October Eurasian SCE and the DJF AO. Confidence intervals for ρ can be calculated using a Fisher-z transformation (Wilks 2011). Using r = −0.40 and a 28-year record, the 95 % confidence interval for ρ is between −0.03 and −0.67. With time, we should get a better sense if the observed correlation is coincidentally strong and will weaken with a longer record, or if model improvements will eventually lead to stronger correlations that are comparable to the observed record. The fact that model runs with realistic prescribed snow cover (e.g., Orsolini and Kvamstø 2009; Allen and Zender 2011) recover the observed relationship better than runs with coupled snow cover offers some indication that the latter may be true.

4.4 Troposphere–stratosphere–troposphere coupling

We now examine how well the CFSv2 runs capture coupling from the troposphere to the stratosphere, and subsequently from the stratosphere back down to the troposphere. As mentioned previously, these couplings may provide a source of memory in the atmosphere at timescales beyond a few weeks, and, as such, may provide a physical mechanism for extending predictability of the wintertime AO. Some of these results are relevant not only to diagnosing relationships between October Eurasian SCE and the wintertime AO, but also to other mechanisms that involve coupling between the stratosphere and the troposphere.

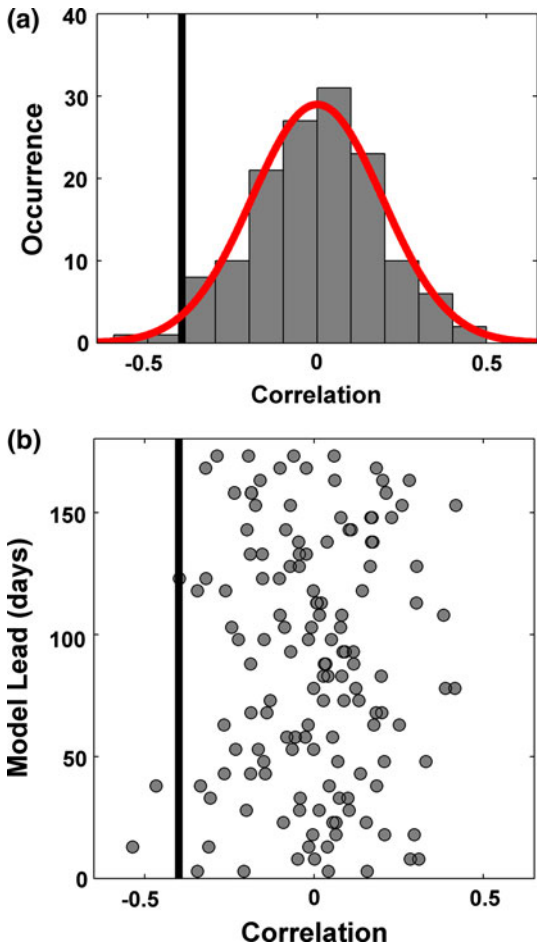

Fig. 8 **a** *Grey bars* show a histogram of correlations values between October Eurasian SCE and the DJF AO index for the 140 CFSv2 runs. The *black vertical line* shows the observed correlation value. The *red curve* shows the null distribution associated with correlations between two independent 28-year time series. **b** Correlation values for each of the 140 CFSv2 ensemble members between October SCE and the normalized DJF AO as a function of model lead-time in days

Figure 9a, b show the relationship between the surface forcing (October Eurasian SCE variability) and vertical wave activity fluxes (WAFz; Plumb 1985) at 40°–80°N for the months of October through February. In the reanalysis, strong spikes in vertical wave activity flux are observed in the troposphere in November and into the stratosphere in December, associated with larger SCE over Eurasia in October. In the CFSv2 model runs, all WAFz anomalies associated with October Eurasian SCE are very weak. A very slight tendency towards positive WAFz at 100 hPa is

observed in January, but the average correlations are a magnitude smaller and occur a month later than in the CFSR.

Figure 9c, d show the relationship between standardized WAFz anomalies in December at 100 hPa, and zonal wind anomalies in October through February. In both the observations and the model, a slowing of the stratospheric polar vortex is observed in the December and January associated with WAFz pulses through the tropopause in December. This relationship is easily explained physically, due to wave breaking events in the stratosphere that slow the vortex (e.g., Matsuno 1971). CFSv2 appears to successfully capture this mechanism.

Figure 10 examines the downward propagation of stratospheric AO signals into the troposphere. The AO in the upper atmosphere is calculated in the same way as the 1,000 hPa AO using an EOF analysis of CFSR December–February monthly geopotential height anomalies. As at 1,000 hPa, CFSv2 AO indices at upper levels are calculated as projections onto the upper level CFSR loading patterns. Figure 10a shows that in CFSR an AO signal at 10 hPa in January is positively correlated with the signal at the surface. In the CFSv2 runs, the coupling extends only down to the tropopause but not below (Fig. 10b). Figure 10c, d are similar, but examine the stratospheric and tropospheric precursor signals associated with surface DJF AO anomalies. Again, the coupling between the surface and the stratosphere is much stronger in the observations than in the model, with the model correlations not extending much above the tropopause. These results suggest a significant shortcoming in the CFSv2 model, given that stratosphere–troposphere coupling is potentially an important source of predictability at both extended range and seasonal timescales (e.g., Orsolini et al. 2011). A similar barrier between the stratosphere and the troposphere is seen in many CMIP-5 models as well (Furtado et al. in revision).

Figure 11 summarizes the stratosphere-troposphere coupling relationships associated with snow cover variability in terms of polar cap height anomalies (i.e., the area-averaged geopotential height anomalies at each pressure level poleward of 60°N). In the observations, positive polar cap height anomalies (i.e. indicative of a negative AO index) associated with October Eurasian SCE extend up to the stratosphere in October (Fig. 11a). This stratospheric AO signal weakens in November but reemerges in December and January, likely in response to anomalous vertical wave propagation and breaking that weakens the stratospheric polar vortex. This signal then influences the surface AO in January and February. By contrast, the CFSv2 ensemble-mean captures a negative AO signal in October, but the signal does not persist into the following winter (Fig. 11b).

We have repeated Figs. 9, 10 and 11 using only the most recent 14 years of the record (not shown). The CFSv2 results for these figures do not depend strongly on whether the full 28-years or only the most recent 14 years are used, suggesting that the higher skill in the second half of the

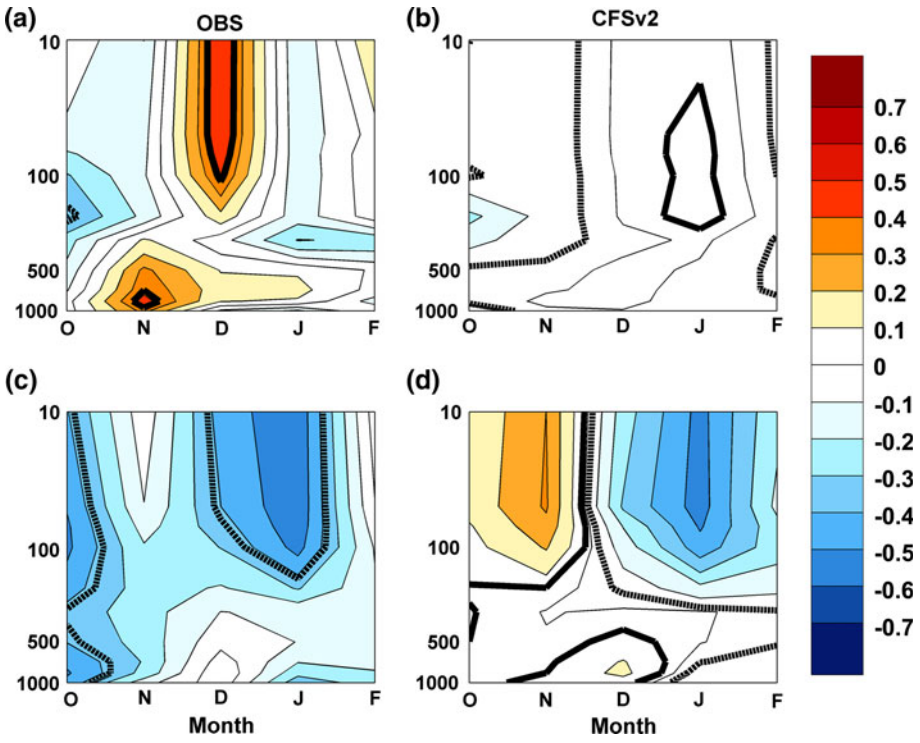

Fig. 9 **a** Correlations between the detrended Rutgers October Eurasian snow cover index and detrended monthly 40–80 N WAFz anomalies. **b** Same as **a** except showing the average of correlations from the 140 CFSv2 hindcast runs. Correlations between the model October snow cover index and the model WAFz anomalies are calculated for each run, separately, then averaged. **c** Correlations between observed standardized Dec 40–80 N WAFz anomalies at 100 mb and zonal wind anomalies at 60 N. **d** Same as **c** except showing the average of correlations from the 140 CFSv2 hindcast runs. In all panels, significant positive (negative) correlations are enclosed with a solid (*dotted*) *black line*

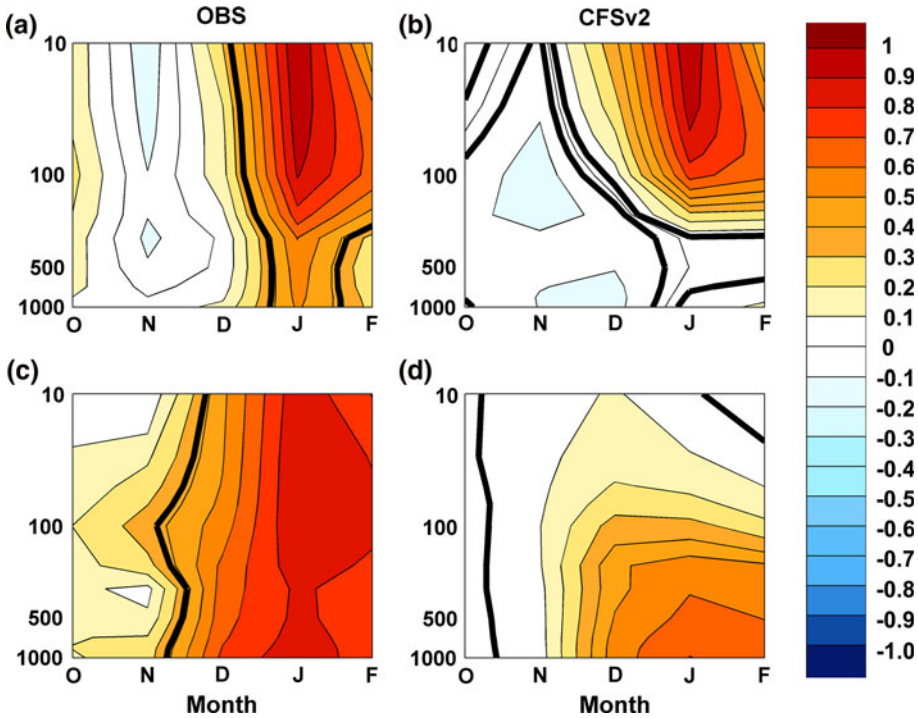

Fig. 10 **a** Correlations between the CFSR January AO index at 10 hPa and monthly AO indices for other months and pressure levels. **b** Same as **a** except showing the average of correlations from the 140 CFSv2 hindcast runs. **c** Correlations between the CFSR DJF AO index at 1,000 hPa and monthly AO indices for other months and pressure levels. **d** Same as **c** except showing the average of correlations from the 140 CFSv2 hindcast runs. In all panels, significant positive (negative) correlations are enclosed with a solid (*dotted*) *black line*

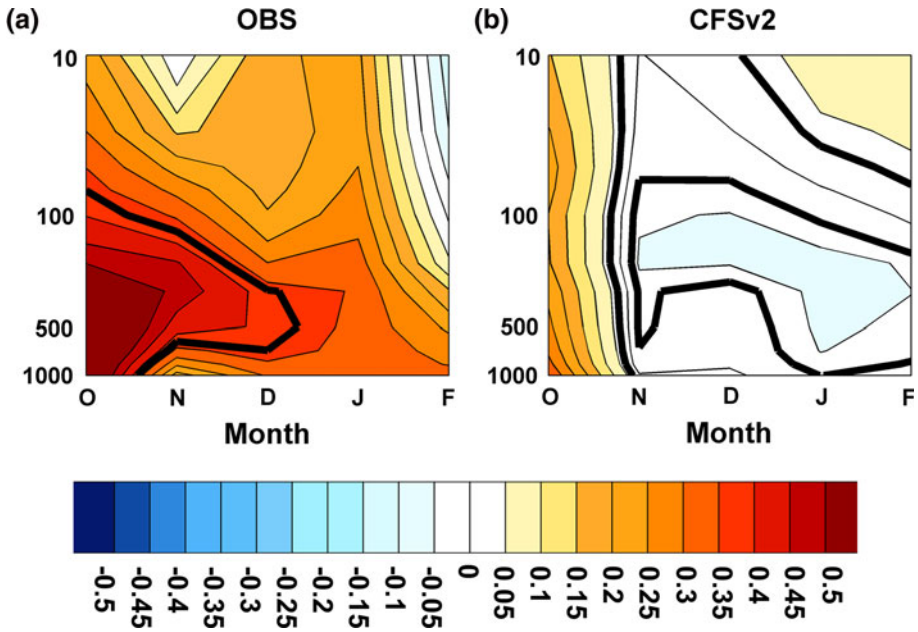

Fig. 11 **a** Correlations between the Rutgers October Eurasian snow cover index and monthly polar cap anomalies. Polar cap anomalies are calculated as the areal average of geopotential height anomalies poleward of 60 N. Significant correlations are enclosed with a *black line*. **b** Same as **a** except showing the average of correlations from the 140 CFSv2 hindcast runs. Correlations between the model October snow cover index and the model polar cap anomalies are calculated for each run separately, and then averaged

record (Fig. 1) cannot be attributed better model representation of troposphere-stratosphere-coupling mechanisms in these years.

5 Discussion and conclusions

This study demonstrates that lagged ensemble-mean forecasts using CFSv2 have small but discernible skill in predicting wintertime AO index at lead-times up to more than 2 months, using a variety of ensemble sizes. While previous studies have also found some skill in dynamical model forecasts of the AO and NAO (e.g., Doblas-Reyes et al. 2003; Müller et al. 2004; Johansson 2007), this is the first to demonstrate skill in an operational model at leads longer than 1 month. The skill was higher if only the most recent half of the record was used.

At leads of 1–3 months, our results suggest that using large ensemble averages of up to 60 or more lagged members may be beneficial if older runs are available.

These results are consistent with the theoretical results of Kumar and Hoerling (2000) who find that, in a "perfect" GCM, the ensemble size needed to achieve the upper limit of predictability increases as the signal to noise ratio in the system decreases, as it does at longer leads. Additional work is also needed to evaluate whether the observed CFSv2 skill is sufficient to be translated into more accurate wintertime climate outlooks such as those issued by CPC.

We applied a simplified dynamic stratification procedure to the ensemble forecasts and found that forecasts using runs with a good representation of October Eurasian SCE do better than forecasts using runs with poor representation of this feature, hinting that links between October Eurasian SCE and the wintertime AO may be responsible for the model skill. However, the improvement is not statistically significant at the 5 % level and run-by-run correlations between October Eurasian SCE and the DJF AO are not distinguishable from a null distribution based on uncorrelated 28-year time series.

Further analysis of this relationship suggests that the CFSv2 model misses several crucial steps in the stratospheric pathway proposed to link October Eurasian SCE to the DJF AO. These results did not change when only the most recent 14 years were used, suggesting that the higher skill seen in recent years cannot be attributed to better representation of stratosphere/troposphere interactions. Model improvements should focus on:

(1) Better representation of the mean climatology and interannual variability of SCE. The model overestimates the total October Eurasian SCE, and its variability in regions of interest (e.g., western Eurasia and south of 60°N) remains poor.
(2) A better relationship between SCE and WAFz. Unlike observations, the CFSv2 model runs do not show a clear lagged response between autumn Eurasian SCE and winter WAFz that impacts the stratospheric circulation (Fig. 9b). However, the model does effectively recover the fundamental dynamical relationship between upwelling tropospheric waves and their impact on the stratospheric polar vortex (Fig. 9d). Therefore, the issue with the model lies in the hypothesized SCE-induced forcing of these waves, not in wave breaking dynamics.
(3) Better representation of the downward propagation of stratospheric anomalies into the troposphere. Despite the effectiveness of the model in capturing the stratospheric response to wave breaking, the downward propagation of the anomalies only exists in the stratosphere and fails to descend into the troposphere, as seen in observations (Fig. 10). This missing aspect of stratosphere-troposphere dynamical coupling has impacts beyond the October Eurasian SCE–DJF AO connection explored in this paper and will likely impact seasonal forecast confidence using the CFSv2 (e.g., Douville 2009; Orsolini et al. 2011; Sigmond et al. 2013). Possible causes may be inaccurate diagnosis of the residual circulation induced by wave breaking at successively lower levels of the stratosphere ('downward control; Haynes et al. 1991) or the lack of tropospheric eddy feedbacks at work in the model (e.g., Robinson 2000; Song and Robinson 2004). The CMIP5 coupled models also suffer from the same lack in downward propagation of stratospheric anomalies into the troposphere (Furtado et al. in revision), and hence future research should focus on this aspect in operational and coupled climate models.

The largest outstanding question raised by this study concerns the source of model skill. If the model is not capturing coupling between the troposphere and the stratosphere, then what mechanism is responsible in the model for translating initial conditions in summer and fall into predictable wintertime climate anomalies? One possibility is that the model has some skill in predicting SST anomalies and that these are forcing wintertime AO anomalies. Future work will focus on understanding alternative mechanisms that may account for skill in the model.

Acknowledgments Work for this project was supported by NOAA Grant #NA10OAR4310163. The observational snow cover dataset was provided by Rutgers University Global Snow Lab. We greatly appreciate helpful editorial comments provided by Dan Collins, Craig Long and two anonymous reviewers.

References

Allen RJ, Zender CS (2010) Effects of continental-scale snow albedo anomalies on the wintertime Arctic oscillation. J Geophys Res 115:D23105. doi:10.1029/2010JD014490

Allen RJ, Zender CS (2011) Forcing of the Arctic Oscillation by Eurasian snow cover. J Clim 24:6528–6539. doi:10.1175/2011JCLI4157.1

Baldwin MP, Dunkerton TJ (1999) Propagation of the Arctic Oscillation from the stratosphere to the troposphere. J Geophys Res 104:30937–30946

Baldwin MP, Dunkerton TJ (2001) Stratospheric harbingers of anomalous weather regimes. Science 294:581–584. doi:10.1126/science.1063315

Buermann W, Anderson B, Tucker CJ, Dickenson RE, Lucht W, Potter CS, Myneni RB (2003) Interannual covariability in Northern Hemisphere air temperatures and greenness associated with El Niño-Southern Oscillation and the Arctic Oscillation. J Geophys Res 108:4396. doi:10.1029/2002JD002630

Chen M, Wang W, Kumar A (2013) Lagged ensembles, forecast configuration, and seasonal predictions. Mon Weather Rev. doi:10.1175/MWR-D-12-00184.1

Cohen J, Barlow M (2005) The NAO, the AO, and global warming: How closely related? J Clim 18:4498–4513. doi:10.1175/JCLI3530.1

Cohen J, Entekhabi D (2001) The influence of snow cover on Northern Hemisphere climate variability. Atmos Ocean 39: 35–53

Cohen J, Fletcher C (2007) Improved skill of Northern Hemisphere winter surface temperature predictions based on land–atmosphere fall anomalies. J Clim 20:4118–4132. doi:10.1175/JCLI4241.1

Cohen J, Jones J (2011a) A new index for more accurate winter predictions. Geophys Res Lett 38:L21701. doi:10.1029/2011GL049626

Cohen J, Jones J (2011b) Tropospheric precursors and stratospheric warmings. J Clim 24:6562–6572. doi:10.1175/2011JCLI4160.1

Cohen J, Barlow M, Kushner PJ, Saito K (2007) Stratosphere–troposphere coupling and links with Eurasian land surface variability. J Clim 20:5335–5343. doi:10.1175/2007JCLI1725.1

Déqué M (1997) Ensemble size for numerical seasonal forecasts. Tellus A 49A:74–86

Derome J, Lin H, Brunet G (2005) Seasonal forecasting with a simple general circulation model: predictive skill in the AO and PNA. J Clim 18:597–609

Deser C, Tomas R, Alexander M, Lawrence D (2010) The seasonal atmospheric response to projected Arctic sea ice loss in the late twenty-first century. J Clim 23:333–351. doi:10.1175/2009JCLI3053.1

Doblas-Reyes FJ, Pavan V, Stephenson DB (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

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

Ek MB, Mitchell KE, Lin Y, Rogers E, Grunmann P, Koren V, Gayno G, Tarpley JD (2003) Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. J Geophys Res 108:8851. doi:10.1029/2002JD003296

Fletcher CG, Kushner PJ (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

Fletcher CG, Hardiman SC, Kushner PJ, Cohen J (2009) The dynamical response to snow cover perturbations in a large ensemble of atmospheric GCM integrations. J Clim 22:1208–1222. doi:10.1175/2008JCLI2505.1

Garfinkel CI, Hartmann DL (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

Garfinkel CI, Hartmann DL (2008) Different ENSO teleconnections and their effects on the stratospheric polar vortex. J Geophys Res 113:D18114. doi:10.1029/2008JD009920

Garfinkel CI, Hartmann DL, Sassi F (2010) Tropospheric precursors of anomalous Northern Hemisphere stratospheric polar vortices. J Clim 23:3282–3299. doi:10.1175/2010JCLI3010.1

Garfinkel CI, Butler AH, Waugh DW, Hurwitz MM, Polvani LM (2012) Why might stratospheric sudden warmings occur with similar frequency in El Niño and La Niña winters? J Geophys Res 117:D19106. doi:10.1029/2012JD017777

Gong G, Entekhabi D, Cohen J (2003) Modeled Northern Hemisphere winter climate response to realistic Siberian snow anomalies. J Clim 16:3917–3931

Griffies SM, Harrison MJ, Pacanowski RC, Rosati A (2004) A technical guide to MOM4. GFDL Ocean Group technical report no 5, p 342

Hardiman SC, Kushner PJ, Cohen J (2008) Investigating the ability of general circulation models to capture the effects of Eurasian snow cover on winter climate. J Geophys Res 113:D21123. doi:10.1029/2008JD010623

Haynes PH, Marks CJ, McIntyre ME , Shepherd TG, Shine KP (1991) On the "downward control" of extratropical diabetic circulations by eddy-induced mean zonal forces. J Atmos Sci 48:651–678

Hu Z–Z, Huang B (2006) On the significance of the relationship between the North Atlantic Oscillation in early winter and Atlantic sea surface temperature anomalies. J Geophys Res 111:D12103. doi:10.1029/2005JD006339

Hurwitz MM, Newman PA, Garfinkel CI (2012) On the influence of North Pacific sea surface temperature on the Arctic winter climate. J Geophys Res 117:D19110. doi:10.1029/2012JD017819

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

Ineson S, Scaife AA, Knight JR, Manners JC, Dunstone NJ, Gray LJ, Haigh JD (2011) Solar forcing of winter climate variability in the Northern Hemisphere. Nat Geosci 4:753–757. doi:10.1038/ngeo1282

Johansson Å (2007) Prediction Skill of the NAO and PNA from Daily to Seasonal Time Scales. J Clim 20:1957–1975. doi:10.1175/JCLI4072.1

Kharin VV, Zwiers FW, Gagnon N (2001) Skill of seasonal hindcasts as a function of the ensemble size. Clim Dyn 17:835–843

Kim H-M, Webster PJ, Curry JA (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

Kopp TJ, Kiess RB (1996) The air force global weather central snow analysis model. In: Preprint 15th conference on weather analysis and forecasting, American Meteorological Society, Norfolk, VA, pp 220–222

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

Kumar A, Hoerling M (2000) Analysis of a conceptual model of seasonal climate variability and implications for seasonal prediction. Bull Am Meteorol Soc 81:255–264

Kumar A, Barnston AG, Hoerling MP (2001) Seasonal predictions, probabilistic verifications, and ensemble size. J Clim 14:1671–1676

Liu J, Curry JA, Wang H, Song M, Horton RM (2012) Impact of declining Arctic sea ice on winter snowfall. Proc Natl Acad Sci 109:4074–4079. doi:10.1073/pnas.1114910109

Lorenz DJ, Hartmann DL (2003) Eddy-zonal flow feedback in the Northern Hemisphere winter. J Clim 16:1212–1227

Lu B-W, Pandolfo L (2011) Nonlinear relation of the Arctic oscillation with the quasi-biennial oscillation. Clim Dyn 36:1491–1504. doi:10.1007/s00382-010-0773-7

Matsuno T (1971) A dynamical model of the stratospheric sudden warming. J Atmos Sci 28:1479–1494

Meehl GA, Hu A, Tebaldi C (2010) Decadal prediction in the Pacific region. J Clim 23:2959–2973. doi:10.1175/2010JCLI3296.1

Müller WA, Appenzeller C, Schär C (2004) 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

Orsolini YJ, Kvamstø NG (2009) Role of Eurasian snow cover in wintertime circulation: decadal simulations forced with satellite observations. J Geophys Res 114:D19108. doi:10.1029/2009JD012253

Orsolini YJ, Kindem IT, Kvamstø NG (2011) 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

Peings Y, Saint-Martin D, Douville H (2012) A numerical sensitivity study of the influence of Siberian snow on the Northern Annular Mode. J Clim 25:592–607. doi:10.1175/JCLI-D-11-00038.1

Plumb RA (1985) On the three-dimensional propagation of stationary waves. J Atmos Sci 42:217–229

Ramsay BH (1998) The interactive multisensor snow and ice mapping system. Hydrol Proc 12:1537–1546

Robinson DA, Dewey KF, Heim RR (1993) Global snow cover monitoring: An update. Bull Am Meteorol Soc 74:1689–1696

Robinson WA (2000) A baroclinic mechanism for the eddy feedback on the zonal index. J Atmos Sci 57:415–422. doi:10.1175/1520-0469(2000)057<0415:ABMFTE>2.0.CO;2

Ruzmaikin A, Feynman J (2002) Solar influence on a major mode of atmospheric variability. J Geophys Res. 107:D144209. doi:10.1029/2001/JD001239

Saha S, Moorthi S, Pan H-L et al (2010) The NCEP climate forecast system reanalysis. Bull Am Meteorol Soc 91:1015–1057. doi:10.1175/2010Bams3001.1

Schubert S, Suarez M, Schemm J-K, Epstein E (1992) Dynamically stratified Monte Carlo forecasting. Mon Weather Rev 120:1077–1088

Sigmond M, Scinocca JF, Kharin VV, Shepherd TG (2013) Enhanced seasonal forecast skill following stratospheric sudden warmings. Nat Geosci 6:98–102. doi:10.1038/ngeo1698

Song Y, Robinson WA (2004) Dynamical mechanisms for stratospheric influences on the troposphere. J Atmos Sci 61:1711–1725

Strong C, Magnusdottir G (2011) Dependence of NAO variability on coupling with sea ice. Clim Dyn 36:1681–1689. doi:10.1007/s00382-010-0752-z

Tang Y, Lin H, Derome J, Tippett MK (2007) A predictability measure applied to seasonal predictions of the Arctic Oscillation. J Clim 20:4733–4750. doi:10.1175/JCLI4276.1

Thompson D, Wallace J (1998) The Arctic Oscillation signature in the wintertime geopotential height and temperature fields. Geophys Res Lett 25:1297–1300

Thompson DWJ, Wallace JM (2001) Regional climate impacts of the Northern Hemisphere annular mode. Science 293:85–89. doi:10.1126/science.1058958

Wang W, Anderson BT, Kaufmann RK, Myneni RB (2004) The relation between the North Atlantic Oscillation and SSTs in the North Atlantic basin. J Clim 17:4752–4759

Wilks DS (2011) Statistical methods in the atmospheric sciences, International Geophysics Series, 3rd edn, vol 100. Academic Press, London, p 676

Yuan X, Wood EF, Luo L, Pan M (2011) A first look at Climate Forecast System version 2 (CFSv2) for hydrological seasonal prediction. Geophys Res Lett 38:L13402. doi:10.1029/2011GL047792