Seasonal prediction of North American wintertime cold extremes in the GFDL SPEAR forecast system

Liwei Jia[1,2*], Thomas L. Delworth[2], Xiaosong Yang[2], William Cooke[2], Nathaniel C. Johnson[2], Colleen McHugh[3] and Feiyu Lu[4]

[1*]University Corporation for Atmospheric Research, Boulder 80307, CO, US.
[2]Geophysical Fluid Dynamics Laboratory, NOAA/OAR, Princeton, 08540, NJ, US.
[3]Science Applications International Corporation, Reston, 20190, VA, US.
[4]Department of Geosciences, Princeton University, Princeton 08540, NJ, US.

*Corresponding author(s). E-mail(s): Liwei.Jia@noaa.gov;
Contributing authors: Tom.Delworth@noaa.gov; Xiaosong.Yang@noaa.gov; William.Cooke@noaa.gov; Nathaniel.Johnson@noaa.gov; Colleen.Mchugh@noaa.gov; Feiyu.Lu@noaa.gov;

Abstract

Skillful prediction of wintertime cold extremes on seasonal time scales is beneficial for multiple sectors. This study demonstrates that North American cold extremes, measured by the frequency of cold days in winter, are predictable several months in advance in the Geophysical Fluid Dynamics Laboratory's SPEAR (Seamless system for Prediction and EArth system Research) seasonal forecast system. Three predictable components of cold extremes over the North American continent are

found to be skillfully predicted on seasonal scales. One is a trend-like component, which shows a continent-wide decrease in the frequency of cold extremes and is primarily attributable to external radiative forcing. This trend-like component is predictable at least 9 months ahead. The second predictable component displays a dipole structure over North America, with negative signs in the northwest and positive signs in the southeast. This dipole component is predictable with significant correlation skill for 2 months and is a response to the central Pacific ENSO (El Niño-Southern Oscillation) as revealed from SPEAR AMIP-style simulations. The third component with the largest loadings over Canada and the northern U.S. shows significant correlations with snow anomalies over mid-to-high latitudes of the North American continent. Predictions using only the three predictable components yield higher/comparable skill relative to the SPEAR raw forecasts.

Keywords: Seasonal, Prediction, Cold Extremes, North America, SPEAR

1 Introduction

Extreme cold temperatures in wintertime pose challenges for multiple sectors such as electrical and water systems, agriculture, as well as human health (Wang et al, 2016). It can cause infrastructure failures and hazardous traveling conditions. During the February 2021 North American cold wave, records of a century old were broken in Oklahoma and Texas, leading to widespread power outages. It caused at least 278 deaths, and damages around $200 billion in the United States (https://en.wikipedia.org/wiki/February_2021_North_American_cold_wave). The severe impacts of wintertime cold extremes indicate that their skillful prediction provides a great benefit to society.

Studies have shown that cold extremes are predictable on synoptic time scales in numerical models (Dai et al, 2021; Sato et al, 2017). However, a single extreme event is hardly predictable beyond synoptic time scales due to the chaotic nature of the climate system. Nevertheless, the statistics of extreme events in a season can be predicted because of their relationship with

slow-varying climate drivers such as sea surface temperatures (SSTs), land conditions and changes in radiative forcing (Sillmann et al, 2017; Hamilton et al, 2012; Luo and Wang, 2017a,b; Jia et al, 2022). A study by Xiang et al (2020) showed that the frequency of extreme cold days in winter is predicted by the European Center for Medium-range Weather Forecasts (ECMWF) model 2-4 weeks in advance over the Northern Hemisphere land area. They found the sources of the predictability are the land-atmosphere coupling and atmospheric internal modes. Forecasting of extremes on subseasonal time scales was also documented in many other studies (Vitart and Robertson, 2018; Vitart et al, 2019; Miller et al, 2021; Becker et al, 2013). A handful of studies showed that the number of extreme daily events in a season is predictable on seasonal time scales (Miller and Wang, 2019; Pepler et al, 2015; Hamilton et al, 2012; Jia et al, 2022; Luo and Wang, 2017a,b). The predictability on seasonal and longer time scales arises primarily from atmosphere boundary conditions such as SSTs, land conditions and radiative forcing.

The present study focuses on predicting cold extremes on seasonal time scales in the GFDL's (Geophysical Fluid Dynamics Laboratory) SPEAR (Seamless system for Prediction and EArth system Research) seasonal forecast system. Unlike previous studies that assess seasonal prediction skill of extremes at a fixed lead time or over an averaged domain, we are able to identify most predictable patterns of cold extremes over North America (north of 23°N) and explicitly quantify their prediction skill. Moreover, potential predictability sources associated with the most predictable patterns of cold extremes are explored.

2 Model and data

2.1 The SPEAR modeling system

SPEAR is the next-generation GFDL modeling system for seasonal to multidecadal prediction and projection (Delworth et al, 2020). It is a coupled ocean-atmosphere-land-sea ice model. The atmosphere and land components are the GFDL AM4-LM4 models (Zhao et al, 2018); the ocean and sea ice components are the MOM6 and SIS2 models (Adcroft et al, 2019). The medium-resolution version of SPEAR has a resolution of approximately 50 km in the atmosphere/land and a coarser resolution of 1 degree in the ocean. SPEAR has been used for GFDL's real-time seasonal forecasts, contributing to the North American Multimodel Ensemble (NMME) project (Kirtman et al, 2014) since 2021 (https://www.cpc.ncep.noaa.gov/products/NMME/). The details of the SPEAR model and the forecast system including initialization and bias correction techniques are documented in Delworth et al (2020) and Lu et al (2020).

2.2 SPEAR hindcasts

A series of SPEAR hindcasts were conducted to assess the prediction skill of North American cold extremes. This study used hindcasts from 1992 to 2019. To create the hindcasts, 15 ensemble members were initialized on the first day of each month and were integrated for 12 months. The initial conditions for the atmospheric, land and sea ice were from a 5-member ensemble of SPEAR restoring simulations in which the atmospheric temperature, moisture and winds were damped back towards values from the Climate Forecast System Reanalysis (CFSR, Saha et al (2010)); the sea surface temperatures (SSTs) were restored to the Optimum Interpolation Sea Surface Temperature (OISST, Reynolds et al (2002)). The oceanic initial conditions of the hindcasts

were from a 30-member ocean analysis, produced by an ocean data assimilation (ODA) run in the coupled SPEAR model. A bias correction scheme called ocean tendency adjustment (OTA) is applied to coupled model seasonal predictions to reduce model drift. A comprehensive assessment of oceanic variability in the ODA can be found in Lu et al (2020). The 15-member ensemble of hindcasts was generated by applying the 5 members of SPEAR restoring simulations to the first 5 members of ODA, the same 5 restoring members to ODA members 6 through 10, and the same 5 restoring members to ODA members 11 through 15. SPEAR has shown significant forecast skill in the prediction of the Niño 3.4 index, surface air temperature over land, midlatitude baroclinic waves, North American winter temperature swing index, Antarctic sea ice, North American summertime heat extremes, Kuroshio Extension and atmospheric rivers over western North America (Lu et al, 2020; Tseng et al, 2021; Bushuk et al, 2021; Zhang et al, 2021; Yang et al, 2022; Jia et al, 2022; Joh et al, 2022).

2.3 SPEAR historical simulations

Thirty members of SPEAR historical simulations were created using initial conditions from widely separated (20 yr) points of the SPEAR control simulation with atmospheric composition fixed at levels representative of the calendar year 1850. Time-varying historical radiative forcings are applied to the period from 1921 to 2014, while projections for the Shared Socioeconomic Pathway 5-8.5 (SSP5-8.5) (Riahi et al, 2017; Kriegler et al, 2017) are applied for years after 2014. In this study, we analyze 15 members of SPEAR historical simulations over 1992-2019 to identify the externally-forced pattern of the frequency of wintertime cold extremes over North America.

2.4 SPEAR AMIP-style simulations

In order to assess the impact of the El Niño-Southern Oscillation (ENSO) on North American winter cold extremes, we conducted a pair of AMIP-style (Atmospheric Model Intercomparison Project) simulations with SPEAR. This pair of experiments is identical to the SPEAR control simulations with atmospheric composition fixed at levels representative of the calendar year 1850, except that the SSTs were replaced with the observed SSTs. The first experiment (called ClimSST) is forced by observed climatological SSTs from HadISST1 (Rayner et al, 2003), where the SST climatology is calculated during 1979-2018. The second experiment (called Nino4SST) is forced by the same climatological SSTs as in the first experiment plus the Niño 4 SST anomalies over the tropical Pacific (20°S-20°N). The Niño 4 SST anomalies were derived as the composite of SSTs when the Niño 4 index (i.e., SST anomalies averaged in the central equatorial Pacific from 5°S - 5°N, 160°E - 150°W) in the DJF season exceeds one standard deviation (Fig.S1). As the only difference between the two simulations is the Niño 4 SSTs over the tropical Pacific, any difference between them represents the response to Niño 4. The Niño 4 SSTs are chosen because as will be shown in section 3 that central Pacific SSTs are identified as one of the predictability sources of North American winter cold extremes. We analyzed 100 years of data from each simulation.

2.5 Verification data

For verification, we use the European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis - ERA5 (Hersbach et al, 2020), including hourly maximum 2-m temperature (Tmax) reanalysis, monthly mean 2-m air temperature, geopotential height at 500 hPa, snow cover and SSTs. The daily Tmax of ERA5 is calculated as the maximum hourly

Tmax of the day. The ERA5 reanalysis data are referred to as the observations hereafter. The observed Niño 4 index was downloaded from https://psl.noaa.gov/gcos_wgsp/Timeseries/Data/nino4.long.anom.data. The observed Pacific/North American (PNA) index was downloaded from NOAA (National Oceanic and Atmospheric Administration) Climate Prediction Center (https://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/month_pna_index2.shtml), which is defined as the sum of 500 hPa height anomalies at (15°-25°N, 180°-140°W) and (45°-60°N, 125°-105°W) minus the sum of 500 hPa height anomalies at (40°-50°N, 180°-140°W) and (25°-35°N, 90°-70°W).

3 Methodology

3.1 Review of Average Predictability Time

The Average Predictability Time (APT) analysis is used to identify predictable components of North American cold extremes. This method determines components that maximize predictability. APT was developed by DelSole and Tippett (2009a,b) and has been widely used in predictability and prediction studies on various time scales from subseasonal to decadal scales (Jia et al, 2015; Jia and DelSole, 2011, 2012; Yang et al, 2015; Wu et al, 2016; Zhang et al, 2017; Xiang et al, 2018; Jia et al, 2022).

To determine the predictable components, one must first define a measure of predictability. In the context of ensemble forecasts initialized at distinct start times, a standard measure of predictability is the signal-to-total ratio, which is calculated as:

$$P(\tau) = 1 - \frac{\sigma_\tau^2}{\sigma_\infty^2}, \tag{1}$$

where σ_τ^2 is the forecast variance at lead time τ and σ_∞^2 is the climatological variance. APT is then defined as twice the sum of $P(\tau)$ over all lead times:

$$APT = 2\sum_{\tau=0}^{\infty}\left(1-\frac{\sigma_\tau^2}{\sigma_\infty^2}\right)\Delta\tau. \tag{2}$$

It yields a time scale that agrees with the usual e-folding time for an exponentially decaying forecast signal. In this study, the sum is from 0 to 9 months for the target season of December-February (DJF), with lead 0 month corresponding to the forecast initialized on December 1st, lead 1 month corresponding to the forecast initialized on November 1st, and so forth. Note that the APT is computed from model forecasts that are independent of the observations.

We then seek the weights $\mathbf{q}$, such that $\mathbf{q}^T\mathbf{x}(\tau, e)$ maximizes APT in (2), where $\mathbf{x}(\tau, e)$ is the state vector specifying the amplitudes of ensemble forecasts at fixed lead time τ, and ensemble member e. The superscript T denotes the transpose operation. It can be shown (DelSole and Tippett, 2009b) that maximizing APT leads to a generalized eigenvalue problem with each eigenvector corresponding to a component:

$$2\sum_{\tau=0}^{\infty}\left(\mathbf{\Sigma}_\infty-\mathbf{\Sigma}_\tau\right)\Delta\tau\mathbf{q} = \lambda\mathbf{\Sigma}_\infty\mathbf{q}. \tag{3}$$

where $\mathbf{\Sigma}_\tau = \mathbf{x}(\tau, e)\mathbf{x}(\tau, e)^T$is the estimated forecast covariance matrix at lead time τ. The eigenvalue λ gives the APT value, and each eigenvector $\mathbf{q}$ corresponds to a component. By ordering eigenvalues and their associated eigenvectors in a decreasing order, the first component maximizes APT, the second component maximizes APT subject to being uncorrelated with the first, and so on. This decomposition based on APT is analogous to empirical orthogonal function (EOF) analysis, but instead of decomposing variance, we decompose predictability. APT is essentially the sum of signal-to-total ratio

over lead times that can be used to measure predictability. The spatial pattern associated with each component can be computed by projecting the component time series $\mathbf{q}^T\mathbf{x}(\tau, e)$ on the original data, which is

$$\mathbf{p} = \mathbf{\Sigma}_\infty \mathbf{q}. \tag{4}$$

When solving the above eigenvalue problem, we project the forecast data onto 15 leading principal components (PCs) and then maximize APT only in the subspace spanned by the 15 PCs. This is done because the number of grid points in the forecast data exceeds the number of samples, which results in singular covariance matrices and an eigenvalue problem that cannot be solved. We have determined that the APT values are insensitive to the number of PCs used.

To determine whether a component is predictable, we tested the significance of the APT value of the component using Monte Carlo methods under the null hypothesis that forecasts are drawn independently from a white noise process (DelSole et al, 2011). We randomly generate a data array of the same dimension as the hindcast data by drawing independent numbers from a normal distribution with zero mean and unit variance. Then the APT method is applied to this data to calculate an ordered sequence of APT values. The procedure is repeated 1000 times to determine the 5% significance level as the APT value at the 95th percentile. A component is potentially predictable if its APT value is large enough to reject the null hypothesis at a 5% significance level. The application of the Monte Carlo method in predictability study can be found in previous studies (Jia et al, 2022; DelSole et al, 2011; Jia and DelSole, 2011; Yang et al, 2015; Jia et al, 2015).

3.2 Signal-to-noise maximizing EOF analysis

The signal-to-noise maximizing EOF (Empirical Orthogonal Function) analysis is a method of extracting the forced variability by maximizing the ratio of signal variance to the noise variance in the ensemble model simulations (Ting et al, 2009; Chang et al, 2000; Venzke et al, 1999). The signal variance is defined as the variance of the ensemble mean and the noise variance is estimated as the variance of the deviations of each ensemble member from the ensemble mean. This technique applies a spatial prewhitening transformation that removes any spatial correlations in the internal atmospheric variability (i.e.,"noise") contained in the ensemble mean. Thus, the spatial covariance in the ensemble average of the simulations is purely due to the forced response. Therefore, the signal-to-noise maximizing EOF analysis is able to isolate the forced variability better than ensemble average.

3.3 Definition of cold extreme

There are multiple ways to define cold extremes. In this study, we follow the definition of Zhang et al (2011) and Johnson et al (2018). A day d is identified as a cold day if its daily maximum temperature (Tmax) anomaly falls below the 10th percentile of day d's Tmax anomaly distribution. In the hindcasts, the Tmax anomaly distribution is from 15 ensemble members during the historical period of 1992-2019. In the observations, because of the small sample size, the distribution is based on the 15 days of Tmax anomaly centered on day d during 1992-2019. An example to illustrate the method of defining cold days is shown in Figure 1a for DJF (February 29 is excluded) of an arbitrary year (1997-98) at (110°W, 30°N) in the observations. The location (110°W, 30°N) was chosen because it showed a large frequency of cold days in the winter season of 1997-98. The daily Tmax below the 10th percentile threshold (red line) is

considered to be a cold day. In this case, there are 27 cold days. According to the 10th percentile threshold of Tmax anomaly, the frequency of cold days, defined as TX10p (in %), in the DJF season for each year and each grid point is then calculated. Fig.1b demonstrates the TX10p for the grid point at (110°W, 30°N), which varies over time. The winter of 1997-98 shows the largest TX10p of 30% (i.e., 27days/90days=30%). We choose the winter of 1997-98 to demonstrate the spatial map of the DJF mean Tmax anomaly, which shows a band of northwest-southeast oriented positive anomalies over most areas of North America, and negative anomalies over south U.S., Mexico and the northmost of North America (Fig.1c). The corresponding spatial structure of TX10p (Fig.1d) demonstrates a low frequency of cold days over most areas of North America and high frequency of cold days in a small area of the southern United States, Mexico and the northmost of Canada.

4 Results

4.1 Pointwise skill of North American TX10p

We first examine SPEAR's ability in predicting the frequency of winter cold extremes. The pointwise rank correlation skill of TX10p over North America for lead 0, 2, 5, 9 months is shown in Fig.2. At lead 0 month (i.e., initialized on December 1st), 71.5% of North American land areas show significant correlation skill of TX10p. The skill drops rapidly. At the 9-months lead (i.e., initialized on March 1st of the previous year), there are only 25.4% of areas that show significant skill. Overall, TX10p is more predictable over high latitudes, which may be partly due to the influence of radiative forcing as will be discussed below. The prediction skill of TX10p is very limited over the midlatitude regions.

4.2 The relationship between the skill of TX10p and the skill of mean winter temperature

Earlier studies found that the skill in predicting temperature extremes is often related to the skill in predicting mean temperature (Johnson et al, 2018; Xiang et al, 2020; Hamilton et al, 2012; Pepler et al, 2015). We show in Fig.3 the anomaly correlation skill of DJF mean 2-m air temperature over North America. The winter mean temperature is more predictable over high latitudes of North America and limited skill is shown over midlatitudes. The lack of prediction skill over midlatitudes has been documented in previous studies (Roy et al, 2020; Johnson et al, 2014). The spatial structures of the mean temperature skill bear great similarities with those of TX10p (Fig.2), suggesting that the prediction skill of the cold extremes is strongly related to the skill of mean temperature. In fact, the TX10p is highly negatively correlated with the DJF mean 2-m air temperature in both the observations and model hindcasts (Fig.S2).

The relationship between the winter mean temperature and TX10p is further demonstrated in the plot of the percentage of grid points in each bin sorted by the correlation skill of TX10p versus the correlation skill of DJF mean temperature (Fig.4). The result includes all leads from 0 to 9 months. A value below the diagonal line implies the mean temperature skill is higher than the skill of TX10p. We see more values below the diagonal line, meaning more grids have higher skill in mean temperature than extreme cold temperature. The correlation between the skill of TX10p and DJF mean temperature is 0.72. This echoes the above result that the skill in predicting the cold extremes is correlated with the skill in predicting the mean air temperature.

4.3 Predictable components of North American TX10p

Having shown the point-by-point skill of winter cold extremes, we now identify large-scale predictable components of North American winter TX10p in model hindcasts, then calculate the prediction skill of the components by correlating the time series of the components in the hindcasts and observations, where the time series in the observations were derived by projecting the ERA5 reanalysis onto the components. Note that a lack of skill at the grid-point scale does not necessarily mean skilless locally. Using the statistical optimization method, APT, we are able to find large-scale components with maximum predictability. Figure 5 shows three predictable components of TX10p which demonstrate significant correlation skill. They explain 11%, 12% and 14% of the total variance, respectively. The rest of the components do not show significant skill, and so they are not discussed.

The first component shows continental-scale negative signs over North America with the largest loadings found in northeastern Canada. The time series of the first component are dominated by an increasing trend in both the observations and model hindcasts (Fig.5b). These results indicate that the frequency of cold extremes decreases with time over most of the North American continent and is likely associated with long-term climate change. The first component is highly predictable with significant correlation skill at all leads from 0 to 9 months.

The second component demonstrates a dipole structure with negative loadings over the northwest of North America and positive loadings over Mexico and part of the southern states of the U.S. from Texas to Florida (Fig.5d). This structure shows similarities to the winter mean temperature teleconnection pattern with ENSO. As will be shown below that the second component of TX10p is ENSO-related. The similar ENSO teleconnection patterns in winter

mean temperature and extreme temperature further reveal the close relation between them. The second component of TX10p shows interannual variability (Fig.5e) and is predictable up to 2 months in advance (Fig.5f). The third component shows positive anomalies over Canada and northern U.S., with largest magnitudes over northwest of Great Lakes (around 50°N). Negative signs are seen over western Alaska. This third component is predictable at lead 0 month (i.e., initialized on December 1st).

4.4 The sources of predictability of North American TX10p

An important question to ask is what the predictability sources of North American cold extremes are. This section explores the potential predictability sources that give rise to prediction skill of the components diagnosed above.

To find the predictability source of the first component, we compute the externally forced variability of TX10p over North American continent from SPEAR historical simulations using the signal-to-noise maximizing EOF analysis. The spatial pattern of the forced variability of TX10p (Fig.6) resembles the pattern of the first predictable component of TX10p (Fig.5a) over most of the high latitudes, although some differences are seen in part of northwestern and southern North America. The high spatial correlation between the two patterns (R=0.81) suggests that the first predictable component is primarily the response to external radiative forcing. The differences between the externally-forced pattern and the pattern of the first component suggest other predictability sources may potentially be responsible for the predictability of the first component as well. We notice that considerable interannual variabilities are shown in the observed time series of the first component in addition to the multidecadal trend (Fig.5b). In fact, the correlation between the detrended

time series of the first component and the Niño4 index is 0.39 (p=0.04), suggesting that ENSO plays a minor role in this trend-like component. This result is consistent with the differences between the externally-forced pattern and the pattern of the first component (Fig.5a and Fig.6), with Fig.5a shows relative cooling in northwest of North America and relative warming in southern North America featuring the ENSO-related pattern (Fig.5d) that will be discussed below.

Figure 7 shows the regression coefficients of global SSTs (in DJF) with the second predictable component of TX10p in the observations and the SPEAR hindcasts initialized on December 1st. The role of SSTs in seasonal climate predictions has been documented in early studies Yu and Zhang (2015); Buchan et al (2014); Ratnam et al (2016); McKinnon et al (2016); Hamilton et al (2012); Johnson et al (2018); Jia et al (2016). The SST structure closely resembles the classic pattern of central Pacific El Niño in both the observations and model hindcasts. The anomaly correlation coefficient between the observed time series of the second component and the Niño 4 index is 0.6. The significant correlation implies that the second predictable component of TX10p is ENSO-related.

Having demonstrated the statistical connection between ENSO and the frequency of North American cold extremes, we further demonstrate their link in SPEAR AMIP-style experiments (i.e., ClimSST and Nino4SST). According to the experimental setup, the differences between experiments Nino4SST and ClimSST are merely the responses to Niño 4 SST anomalies. Figures 8a-b show the climatological differences of TX10p and 500 hPa geopotential height between Nino4SST and ClimSST. As expected, a northwestern-southeastern oriented dipole structure in TX10p, resembling the structure of the second predictable component of TX10p, is evident. The negative values are located

over northwestern North America but the positive values are over southeastern North America. The climatological difference in 500 hPa geopotential height shows a wave-train pattern (Fig.8b). The high-pressure system located over northwestern North America corresponds well to the negative TX10p there, and the low-pressure system over the southeastern U.S. corresponds to more cold extremes over southeastern U.S. and Mexico. A TX10p pattern showing similarities to the simulated ENSO-related TX10p pattern (Fig.8a) is seen in the observed composite map of TX10p when the Niño 4 index exceeds one standard deviation (Fig.8c), albeit the observed pattern shows large negative anomalies over the northeastern U.S. SPEAR overestimates the cold extremes over the southeastern U.S. The observed composite map of the 500 hPa geopotential height anomaly also shows a wave-train pattern as in the SPEAR experiments, although the maximum high-pressure system over North America is shifted westward in SPEAR relative to observations. Such bias in ENSO-related teleconnections is also seen in other SPEAR simulations (Delworth et al, 2020).

In order to reveal predictability sources of the third predictable component of TX10p, we examine the land conditions. Figure 9 shows the correlations between the snow cover (snow water) in observations (model hindcasts) and the time series of the third component. Snow anomalies over land serve as an important source for seasonal prediction of cold extremes (Luo and Wang, 2017a,b; Cohen et al, 2013). The third component is significantly correlated with the DJF mean snow cover anomalies over the northeastern U.S. and northwest of the Great Lakes (Fig.9a). A similar correlation pattern (Fig.S3) is also found for the correlations between the time series of the third component and snow cover anomalies on December 1st (i.e., initial conditions for DJF forecast). The regression pattern of 500 hPa geopotential height anomalies shows

a strong high-pressure system over the North Pacific near the Aleutian Islands (Fig.9c). A low-pressure system is evident over northeastern North America. This atmospheric circulation favors the polar cold air masses near Alaska moving southeastward to mid-to-high latitudes (40°-60°N) of the North American continent. In SPEAR hindcasts, the column-accumulated snow water positively correlates with the third component in mid-to-high latitudes of North America, especially over the northeastern U.S. (Fig.9b). This correlation pattern in SPEAR shows some similarities to the observed pattern of snow cover anomalies, although some differences are also found. The differences may be partly attributable to the different variables used in the observations and model hindcasts. As in the observations, the 500 hPa geopotential height in the model has a high-pressure system over the North Pacific but a low-pressure system over the northern North American continent (Fig.9d). The above results suggest that snow cover initial conditions and land-atmosphere interactions contribute to the prediction skill of the third component of TX10p. Note that the 500 hPa circulation patterns show some similarities with the negative phase of the PNA pattern. The observed time series of the third component is significantly correlated with the observed PNA index with an anomaly correlation of -0.48 (p=0.01), which implies the PNA could partly contribute to the prediction skill of the third component.

4.5 Predictions of TX10p using predictable components

Aiming to optimize prediction skill of TX10p and demonstrate the potential applications for advancing predictions of North American cold extremes, we reconstruct the predictions of TX10p using the three predictable components identified above by multiplying the predictable patterns with their associated time series. As the reconstructed predictions filter out unpredictable noise, they

are expected to show higher skill than model raw predictions without filtering out unpredictable components. Figure 10 demonstrates the correlations skill (averaged over leads 0 to 9 months) of TX10p from SPEAR raw predictions, reconstructed predictions using the three predictable components, and their skill differences. On average, similar structures are shown for the skill of the raw and reconstructed predictions (Figs.10a-b), which demonstrate the largest skill in high latitudes. However, the reconstructed predictions show higher skill over southern and central Canada. In terms of the percentage of areas with significant correlation skill over the North American continent (Fig.10d), the reconstructed predictions have more areas with significant skill than the raw predictions for leads from 1 to 9 months. As for the lead 0 forecasts (i.e., initialized on December 1st.), the raw prediction skill is higher than the reconstructed skill primarily because of the large contributions from initial conditions for the first lead. As the lead time increases, the contributions from boundary conditions such as radiative forcing and SSTs dominate. Overall, predictions using three predictable components yield higher or comparable skill in TX10p.

5 Summary and discussion

This study assesses the prediction skill of North American wintertime cold extremes on seasonal time scales. A statistical optimization technique – APT is applied to identify large-scale predictable components of cold extremes in the GFDL SPEAR seasonal forecast system, where the cold extreme is measured by the frequency of cold days (i.e., TX10p) during the DJF season. The predictability sources of cold extremes are explored as well.

We identified three components of TX10p over North America that are skillfully predictable on seasonal time scales. One is a trend-like component,

which shows a continental scale decrease in the frequency of cold extremes. The trend-like component is highly predictable at least 9 months ahead in SPEAR hindcasts. The external radiative forcing is the primary source of predictability for the trend-like component. The second is an ENSO-related component, showing a southeast-northwest dipole structure over North American land area. This component is predictable with significant correlation skill for 2 months. The link between North American cold extremes and ENSO is further confirmed in SPEAR AMIP-style simulations. A TX10p dipole pattern reminiscent of the ENSO-related component diagnosed from SPEAR hindcasts is revealed as the response to Niño 4 SST anomalies. A similar TX10p pattern appears in the observed composite map of TX10p corresponding with a positive Niño 4 index. The third component with largest loadings over Canada and the northern U.S. shows a relationship with snow anomalies there. Predictions using only the three predictable components lead to higher or comparable skill in TX10p.

Overall, we diagnosed three components of North American wintertime cold extremes in SPEAR that are skillfully predictable on seasonal time scales, and attributed the sources of such predictability as radiative forcing, central Pacific El Niño and snow cover anomalies. In addition, we demonstrated the potential applications for improving seasonal predictions of North American cold extremes by reconstructing predictions using three predictable components. The results that external radiative forcing, ENSO and snow anomalies contribute to the prediction skill of North American wintertime cold extremes do not exclude other potential predictability sources of cold extremes such as the stratosphere (Thompson et al. 2002; Butler et al. 2019; Jia et al. 2017; Cohen et al. 2021). In the SPEAR seasonal forecast system, we do not find a significant association between the stratospheric polar vortex and cold extremes in

the DJF season. Possible reasons why components due to other predictability sources do not appear in the analysis are: (1) predictability sources or the connections between the drivers and North American cold extremes are not well represented in the model due to model biases; (2) components due to other predictability sources are not predictable on seasonal time scales. A study by Jia et al. (2017) showed that the predictability of northern extratropical surface temperature driven by the stratosphere in DJF is not well predicted in GFDL's FLOR (Forecast-Oriented Low Ocean Resolution) forecast system because of the model's bias in the stratosphere. The SPEAR's performance in simulating the stratosphere and its link to the cold extremes could be explored in future studies.

Predictions of cold extremes benefit from accurate model representations of the predictability sources such as the tropical Pacific SSTs, the radiatively forced temperature trend and the snow anomalies in mid-to-high latitudes of North America. However, state-of-the-art seasonal forecast systems still show difficulties in representing many aspects of the climate system. Studies have shown that the NMME seasonal forecast models demonstrate overly strong temperature trends over much of the United States (Becker et al, 2022), and more positive SST trends in the central and eastern equatorial Pacific Ocean than those observed during 1982-2020 (L'Heureux et al, 2022). The SPEAR seasonal forecast system also shows errors in predicting linear trends of cold extremes over North America during the period of 1992-2019. Such errors may be pertinent to cold extremes prediction in real-time. Future work could be conducted to understand the error source of the model in representing the trend, aiming to reduce the trend errors and to improve forecast skill of cold extremes. Also, note that skill assessed from hindcasts is likely overestimates of real forecast skill because the hindcasts are informed by observations over

the period assessed that would not be available to real forecasts. Changes in model biases from historical period to forecast period can lead to differences between hindcast and forecast skill (Risbey et al, 2021).

Acknowledgments. The authors thank Drs. Hiroyuki Murakami and Bor-Ting Jong for helpful reviews of an earlier draft of this research, and two reviewers for constructive comments that helped to improve this manuscript. This research is supported by NOAA's Science Collaboration Program and administered by UCAR's Cooperative Programs for the Advancement of Earth System Science (CPAESS) under award #NA21OAR4310383

Data Availability Statements. The ERA5 reanalysis datasets are available at the Copernicus Climate Change Service Climate Data Store: https://cds.climate.copernicus.eu. The observed Niño 4 index is downloadable from https://psl.noaa.gov/gcos_wgsp/Timeseries/Data/nino4.long.anom.data. The SPEAR seasonal hindcast data are available at the NMME website (https://www.cpc.ncep.noaa.gov/products/NMME). The SPEAR historical simulations are available at GFDL (https://www.gfdl.noaa.gov/spear_large_ensembles/). Specific variables (if not publicly available) from SPEAR hindcasts/historical simulations are available from the corresponding author upon reasonable request.

Supplementary information. Supplementary information is provided in a separate document.

References

Adcroft A, Anderson W, Balaji V, et al (2019) The GFDL global ocean and sea ice model OM4.0: Model description and simulation features. Journal

of Advances in Modeling Earth Systems 11(10):3167–3211. https://doi.org/https://doi.org/10.1029/2019MS0017268

Becker EJ, van den Dool H, Malaquias P (2013) Short-term climate extremes: Prediction skill and predictability. Journal of Climate 26:512–531. https://doi.org/https://doi.org/10.1175/JCLI-D-12-00177.1

Becker EJ, Kirtman BP, L'Heureux ML, et al (2022) A decade of the North American Multimodel Ensemble (NMME): Research, application, and future directions. Bull Am Meteorol Soc 103:E973–E995. https://doi.org/https://doi.org/10.1175/BAMS-D-20-0327.1

Buchan J, Hirschi JJM, Blaker AT, et al (2014) North Atlantic SST anomalies and the cold North European weather events of winter 2009/10 and december 2010. Monthly Weather Review 142:922–932. https://doi.org/https://doi.org/10.1175/MWR-D-13-00104.1

Bushuk M, Winton M, Haumann FA, et al (2021) Seasonal prediction and predictability of regional Antarctic sea ice. J Climate 34:6207–6233

Chang P, Saravanan R, Ji L, et al (2000) The effect of local sea surface temperatures on atmospheric circulation over the tropical Atlantic sector. J Climate 13:2195–2216

Cohen J, Jones J, Furtado J, et al (2013) Warm arctic, cold continents: A common pattern related to arctic sea ice melt, snow advance, and extreme winter weather. Oceanography 26(4). https://doi.org/http://dx.doi.org/10.5670/oceanog.2013.70

Dai G, Li C, Han Z (2021) The Nature and Predictability of the East Asian Extreme Cold Events of 2020/21. Adv Atmos Sci https://doi.org/https://

doi.org/10.1007/s00376-021-1057-3

DelSole T, Tippett MK (2009a) Average predictability time. Part I: Theory. J Atmos Sci 66:1172–1187

DelSole T, Tippett MK (2009b) Average predictability time. Part II: Seamless diagnoses of predictability on multiple time scales. J Atmos Sci 66:1188–1204

DelSole T, Tippett MK, Shukla J (2011) A significant component of unforced multidecadal variability in the recent acceleration of global warming. J Climate 24:909–926

Delworth T, Cooke WF, Adcroft A, et al (2020) The next generation GFDL modeling system for seasonal to multidecadal prediction and projection. J of Advances in Modeling Earth Systems p e2019MS001895. https://doi.org/https://doi.org/10.1029/2019MS001895

Hamilton E, R.Eade, J.Graham R, et al (2012) Forecasting the number of extreme daily events on seasonal timescales. J GeophysRes 117:D03,114. https://doi.org/doi:10.1029/2011JD016541

Hersbach H, Bell B, Berrisford P, et al (2020) The ERA5 global reanalysis. Quar J Royal Meteorol Soc 146:1999–2049. https://doi.org/https://doi.org/10.1002/qj.3803

Jia L, DelSole T (2011) Diagnosis of multiyear predictability on continental scales. J Climate 24:5108–5124

Jia L, DelSole T (2012) Multi-year predictability of temperature and precipitation in multiple climate models. Geophys Res Lett 39. https://doi.org/doi:10.1029/2012GL052778

Jia L, Yang X, Vecchi GA, et al (2015) Improved seasonal prediction of temperature and precipitation over land in a high-resolution GFDL climate model. J Climate 28:2044–2062. https://doi.org/doi:http://dx.doi.org/10.1175/JCLI-D-14-00112.1

Jia L, Vecchi GA, Yang X, et al (2016) The roles of radiative forcing, sea surface temperatures, and atmospheric and land initial conditions in U.S. summer warming episodes. J Climate 29:4121–4135

Jia L, Delworth TL, Kapnick S, et al (2022) Skillful seasonal prediction of North American summertime heat extremes. Journal of Climate, early release

Joh Y, Delworth TL, Wittenberg AT, et al (2022) Seasonal-to-decadal variability and prediction of the Kuroshio Extension in the GFDL coupled ensemble reanalysis and forecasting system. Journal of Climate 35:3515–3535. https://doi.org/DOI:10.1175/JCLI-D-21-0471.1

Johnson NC, Collins DC, Feldstein SB, et al (2014) Skillful wintertime North American temperature forecasts out to 4 weeks based on the state of ENSO and the MJO. Weather and Forecasting 29:23–38. https://doi.org/https://doi.org/10.1175/WAF-D-13-00102.1

Johnson NC, Xie SP, Kosaka Y, et al (2018) Increasing occurrence of cold and warm extremes during the recent global warming slowdown. Nat Commun 9:1724

Kirtman BP, Min D, Infanti JM, et al (2014) The North American Multimodel Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction. BAMS 95:585–601. https://doi.

org/doi:http://dx.doi.org/10.1175/BAMS-D-12-00050.1

Kriegler E, Bauer N, Popp A, et al (2017) Fossil-fueled development (ssp5): An energy and resource intensive scenario for the 21st century. Global Environmental Change 42:297–315. https://doi.org/https://doi.org/10.1016/j.gloenvcha.2016.05.015, URL https://www.sciencedirect.com/science/article/pii/S0959378016300711

L'Heureux M, Tippett M, Wang W (2022) Prediction challenges from errors in tropical Pacific sea surface temperature trends. Front Clim 4:837,483. https://doi.org/doi:10.3389/fclim.2022.837483

Lu F, Harrison MJ, Rosati A, et al (2020) GFDL's SPEAR Seasonal Prediction System: Initialization and Ocean Tendency Adjustment (OTA) for Coupled Model Predictions. Journal of Advances in Modeling Earth Systems 12:e2020MS002,149. https://doi.org/https://doi.org/10.1029/2020MS002149

Luo X, Wang B (2017a) How predictable is the winter extremely cold days over temperate East Asia? Clim Dyn 48(7-8):2557–2568. https://doi.org/https://doi.org/10.1007/s00382-016-3222-4

Luo X, Wang B (2017b) Predictability and prediction of the total number of winter extremely cold days over China. Clim Dyn https://doi.org/DOI10.1007/s00382-017-3720-z

McKinnon K, Rhines A, Tingley M, et al (2016) Long-lead predictions of eastern United States hot days from pacific sea surface temperatures. Nature Geosci 9:389–394

Miller D, Wang Z, Li B, et al (2021) Skillful subseasonal prediction of U.S. extreme warm days and standardized precipitation index in boreal summer. Journal of Climate 34:5887–5898, https://doi.org/10.1175/JCLI–D–20–0878.1

Miller DE, Wang Z (2019) Skillful seasonal prediction of Eurasian winter blocking and extreme temperature frequency. Geophys Res Lett 46:11,530–11,538. https://doi.org/https://doi.org/10.1029/2019GL085035

Pepler AS, Díaz LB, Prodhomme C, et al (2015) The ability of a multi-model seasonal forecasting ensemble to forecast the frequency of warm, cold and wet extremes. Weather and Climate Extremes 9:68–77

Ratnam JV, Behera SK, Annamalai H, et al (2016) ENSO's far reaching connection to Indian cold waves. Scientific Reports 6:37,657. https://doi.org/10.1038/srep37657

Rayner NA, Parker DE, Horton EB, et al (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. J Geo Res 108:NO. D14, 4407. https://doi.org/doi:10.1029/2002JD002670

Reynolds RW, Rayner NA, Smith TM, et al (2002) An improved in situ and satellite SST analysis for climate. J Climate 15:1609–1625

Riahi K, van Vuuren DP, Kriegler E, et al (2017) The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: An overview. Global Environmental Change 42:153–168. https://doi.org/https://doi.org/10.1016/j.gloenvcha.2016.05.009

Risbey JS, Squire DT, Black AS, et al (2021) Standard assessments of climate forecast skill can be misleading. Nature Communications 12:E973–E995. https://doi.org/10.1038/s41467-021-23771-z

Roy T, He X, Lin P, et al (2020) Global evaluation of seasonal precipitation and temperature forecasts from NMME. J Hydrometeorology 21:2473–2486

Saha S, Adcroft, Adcroft, et al (2010) The NCEP Climate Forecast System Reanalysis. Bull Amer Meteor Soc 91:1015–1057

Sato K, Inoue J, Yamazaki A, et al (2017) Improved forecasts of winter weather extremes over midlatitudes with extra Arctic observations. J Geophys Res Oceans 122:775–787. https://doi.org/doi:10.1002/2016JC012197

Sillmann J, Thorarinsdottir T, Keenlyside N, et al (2017) Understanding, modeling and predicting weather and climate extremes: Challenges and opportunities. Weather and Climate Extremes 18:65–74. https://doi.org/https://doi.org/10.1016/j.wace.2017.10.003, URL https://www.sciencedirect.com/science/article/pii/S2212094717300440

Ting M, Kushnir Y, Seager R, et al (2009) Forced and internal twentieth-century SST trends in the North Atlantic. J Climate 22:1469–1481

Tseng KC, Johnson NC, Kapnick SB, et al (2021) Are multiseasonal forecasts of atmospheric rivers possible? Geophys Res Lett, 48 p e2021GL094000

Venzke S, Allen MR, Sutton RT, et al (1999) The atmospheric response over the North Atlantic to decadal changes in sea surface temperature. J Climate 12:2562–2584

Vitart F, Robertson A (2018) The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. npj Clim Atmos Sci 1:3. https://doi.org/https://doi.org/10.1038/s41612-018-0013-0

Vitart F, Cunningham C, DeFlorio M, et al (2019) Sub-seasonal to Seasonal Prediction of Weather Extremes, Sub-Seasonal to Seasonal Prediction. Elsevier

Wang L, Liu T, Hu M (2016) The impact of cold spells on mortality and effect modification by cold spell characteristics. Sci Rep p 38380. https://doi.org/https://doi.org/10.1038/srep38380

Wu Y, M.Latif, Park W (2016) Multiyear predictability of Northern Hemisphere surface air temperature in the Kiel Climate Model. Clim Dyn 47:793–804. https://doi.org/https://doi.org/10.1007/s00382-015-2871-z

Xiang B, Lin SJ, Zhao M, et al (2018) Subseasonal week 3-5 surface air temperature prediction during boreal wintertime in a GFDL model. Geophys Res Lett 46:416–425. https://doi.org/https://doi.org/10.1029/2018GL081314

Xiang B, Sun Y, Chen J, et al (2020) Subseasonal prediction of land cold extremes in boreal wintertime. Journal of Geophysical Research: Atmospheres 125. https://doi.org/https://doi.org/10.1029/2020JD032670

Yang X, Vecchi GA, Gudgel RG, et al (2015) Seasonal predictability of extratropical storm tracks in GFDL's high-resolution climate prediction model. J Climate 28:3592–3611. https://doi.org/doi:http://dx.doi.org/10.1175/JCLI-D-14-00517.1

Yang X, Delworth TL, Jia L, et al (2022) On the seasonal prediction and predictability of winter surface Temperature Swing Index over North America.

Front Clim p 4:972119. https://doi.org/doi:10.3389/fclim.2022.972119

Yu B, Zhang X (2015) A physical analysis of the severe 2013/2014 cold winter in North America. Journal of Geophysical Research:Atmospheres 120:10,149–10,165. https://doi.org/10.1002/2015JD023116

Zhang G, Murakami H, Cooke WF, et al (2021) Seasonal predictability of baroclinic wave activity. npj Climate and Atmospheric Science 4. https://doi.org/https://doi.org/10.1038/s41612-021-00209-3

Zhang L, Delworth TL, Jia L (2017) Diagnosis of decadal predictability of southern ocean sea surface temperature in the gfdl cm2.1 model. Journal of Climate 30:6309–6328. https://doi.org/DOI:https://doi.org/10.1175/JCLI-D-16-0537.1

Zhang X, Alexander L, Hegerl G, et al (2011) Indices for monitoring changes in extremes based on daily temperature and precipitation data. Wiley Interdisciplinary Reviews: Climate Change 2(6):851–870

Zhao M, Golaz JC, Held IM, et al (2018) The GFDL Global atmosphere and land model AM4.0/LM4.0: 1. Simulation characteristics with prescribed SSTs. Journal of Advances in Modeling Earth Systems 10(3):691–734. https://doi.org/https://doi.org/10.1002/2017MS001208

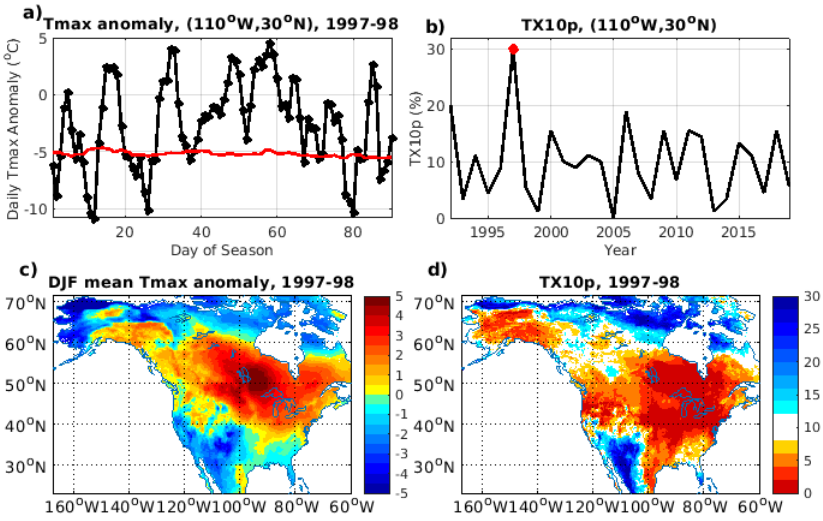

Fig. 1 (a) The observed daily maximum temperature (Tmax, °C) anomaly at (110°W, 30°N) from December 1st, 1997 to February 28th, 1998. The red line indicates the daily 10th percentile threshold. (b) The year-to-year variation of TX10p at (110°W, 30°N) in DJF from 1992-93 to 2019-20. Units: % (c) The DJF mean Tmax anomalies (°C) over North America in the winter of 1997-98. (d) The TX10p (%) for the winter of 1997-98. Data used in this figure are from ERA5 reanalysis.

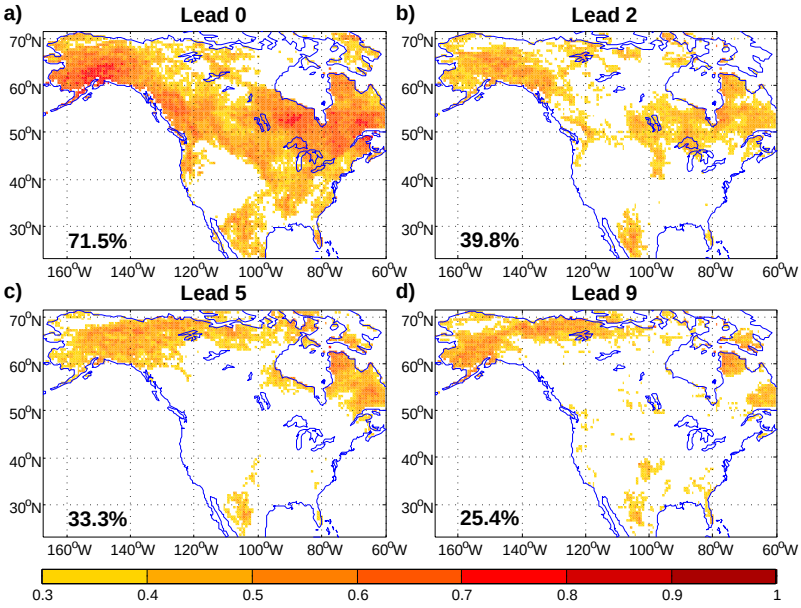

Fig. 2 The pointwise rank correlation skill for the frequency of North American winter cold extremes (i.e., TX10p) at lead (a) 0, (b) 2, (c) 5, (d) 9 months. The value in each panel is the percentage of significant points over North American land area. The statistical significance of the skill is based on the one-tailed Student's t-test at 5% significance level. Insignificant points are shown in white. The verification data used here are from ERA5.

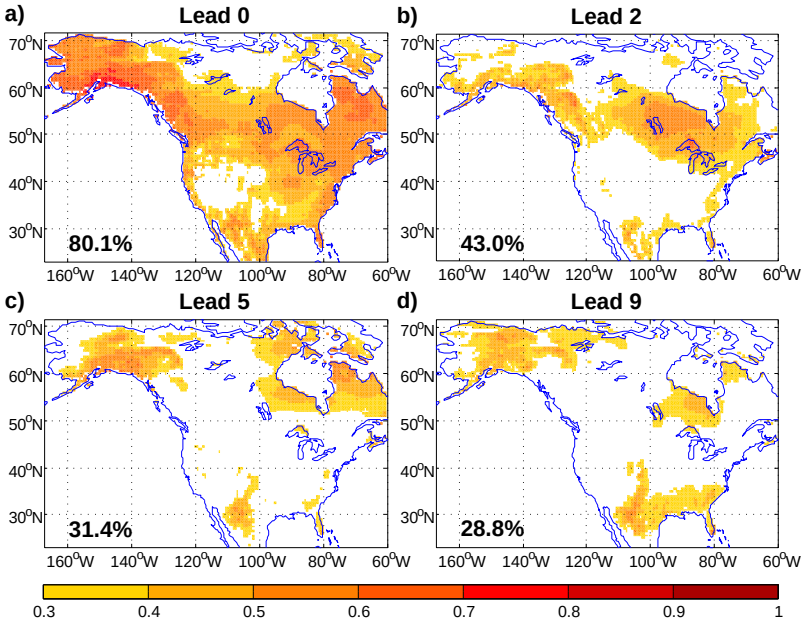

Fig. 3 The pointwise anomaly correlation skill of DJF mean 2-m air temperature over North American land area at lead (a) 0, (b) 2, (c) 5, (d) 9 months. The value in each panel is the percentage of significant points over North American land area. The statistical significance of the skill is based on the one-tailed Student's t-test at 5% significance level. Insignificant points are shown in white.

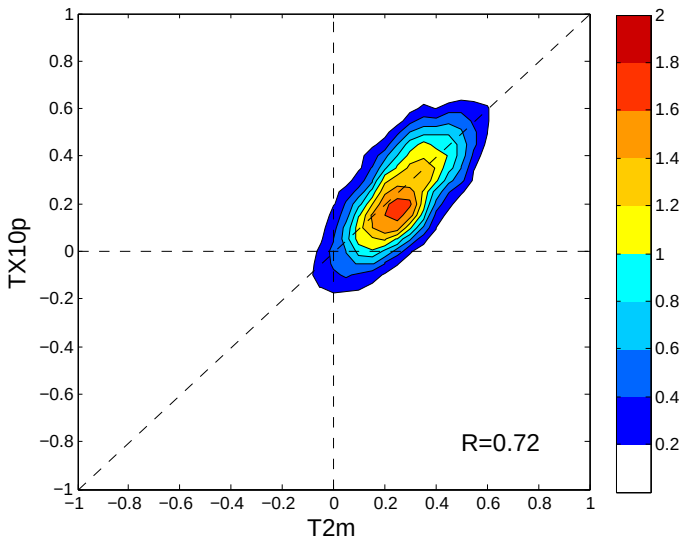

Fig. 4 The percentage of grid points in each bin sorted by the rank correlation skill of TX10p (y-axis) versus anomaly correlation skill of DJF mean 2-m air temperature (x-axis) over North American land area, where each bin has a width of 0.05, for all lead months from 0 to 9 months. The correlation between the two samples is 0.72.

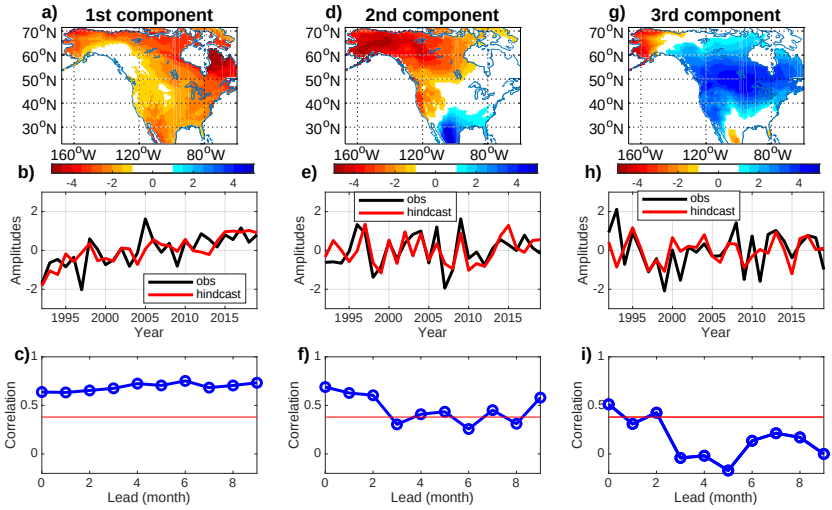

Fig. 5 (a,d,g) The spatial patterns (in %), (b,e,h) time series and (c,f,i) rank correlation skill of three predictable components of North American winter (DJF) TX10p. The hindcasted time series in b), e) and h) (red lines) are the ensemble mean of hindcasts initialized on December 1st (i.e., lead 0 month). The horizontal lines in c), f) and i) indicate the 5% significance level.

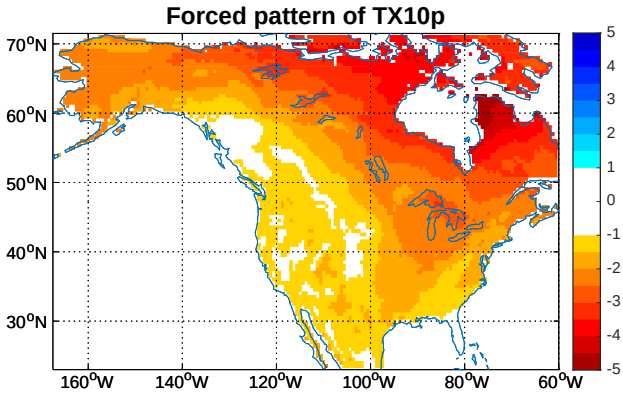

Fig. 6 The externally-forced pattern of TX10p derived from SPEAR historical simulations using signal-to-noise maximizing EOF analysis. Units in %.

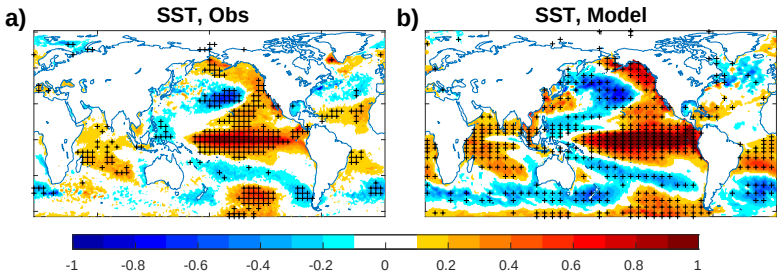

Fig. 7 Regression maps of DJF mean SST (Units: K) on the time series of the second predictable component of TX10p in the (a) observations and (b) SPEAR hindcasts initialized on December 1st. Stippling indicates the regression coefficients are statistically significant at the 5% level.

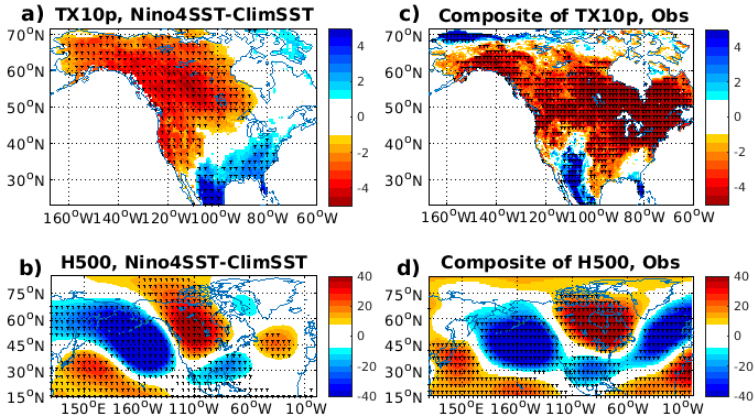

Fig. 8 The climatological differences between the Niño4SST and ClimSST experiments in (a) TX10p (Units: %) and (b) 500 hPa geopotential height (Units: m); and the observed composite of (c) TX10p and (d) 500 hPa geopotential height anomalies when the Niño 4 index exceeds one standard deviation. Stippling indicates the fields are statistically significant at the 5% level.

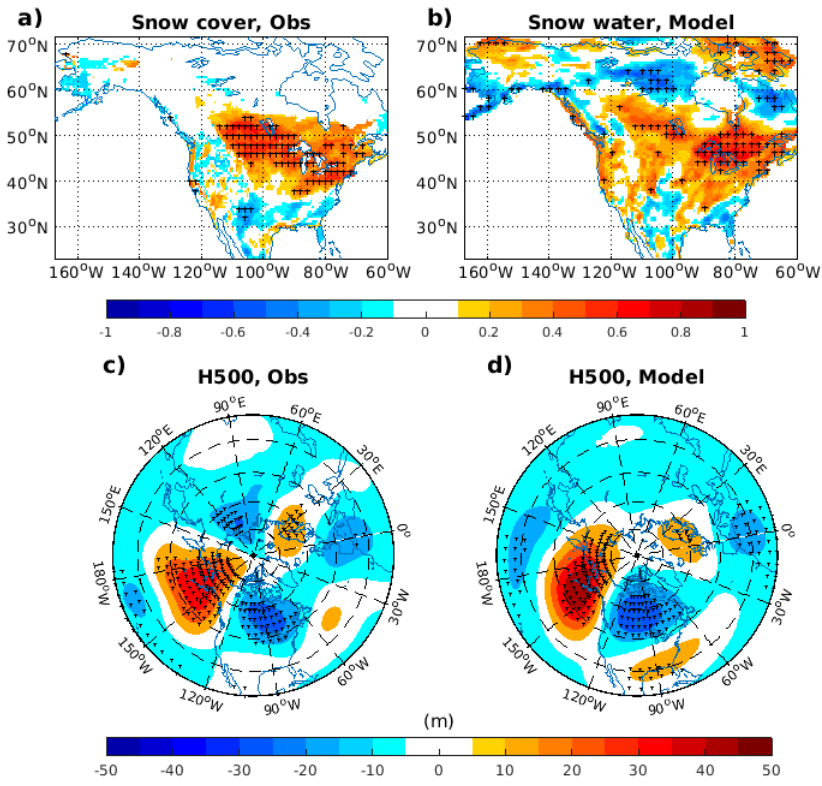

Fig. 9 (Top) Correlation coefficients between the time series of the third predictable component of DJF TX10p and DJF mean (a) snow cover anomalies in the observations and (b) column-integrated snow water in SPEAR hindcasts initialized on December 1st. (Bottom) Regression maps of 500 hPa geopotential height (Units: m) on the time series of the third predictable component of TX10p in (c) observations and (d) model hindcasts. Stippling indicates the correlation/regression coefficients are statistically significant at the 5% level.

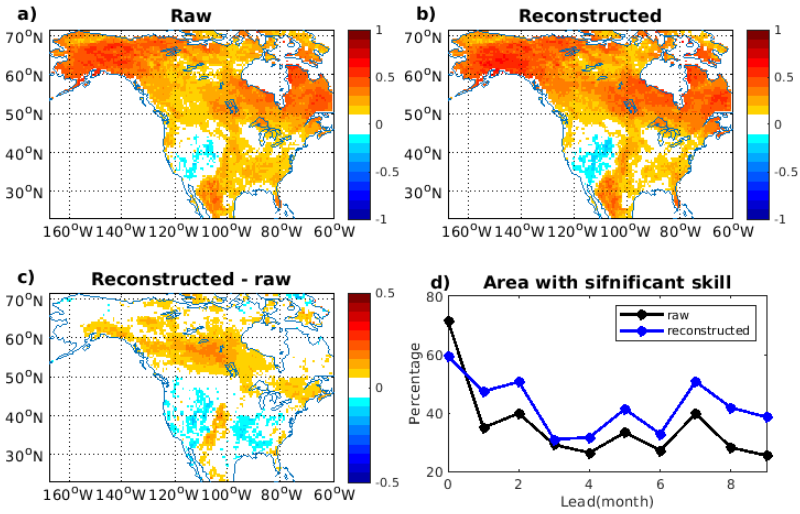

Fig. 10 The rank correlations skill of TX10p in (a) raw predictions directly from SPEAR, (b) reconstructed predictions from three predictable components shown in Fig.5, (c) the skill from reconstructed predictions minus the skill from raw predictions, (d) the percentage of North American land area with significant rank correlation skill at the 5% significance level based on one-tailed Student's t-test in model raw predictions (black) and reconstructed predictions (blue) as a function of lead time. The correlations in (a), (b) and (c) are averaged over lead times from 0 to 9 months.