Evaluation of IPCC Models' Performance in Simulating Late-Twentieth-Century Climatologies and Weather Patterns over North America

VALENTINA RADIĆ AND GARRY K. C. CLARKE

Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, Canada

(Manuscript received 29 December 2010, in final form 12 May 2011)

ABSTRACT

The authors analyze the performance of 22 Intergovernmental Panel on Climate Change (IPCC) global climate models (GCMs) over all of North America and its western subregion using several different evaluation metrics. They assess the model skill in simulating climatologies of several climate variables and the skill in simulating the daily synoptic patterns. The evaluation is performed by comparing the model output with the North American Regional Reanalysis (NARR) over the period 1980–99. One set of metrics, based on root-mean-square errors and variance ratios, compares modeled versus the NARR mean annual cycle and interannual variability. Based on these measures the three top performing models are the ECHAM5–Max Planck Institute Ocean Model (MPI-OM), the third climate configuration of the Met Office Unified Model (HadCM3), and the Canadian Centre for Climate Modelling and Analysis (CCCma) Coupled General Circulation Model, version 3.1 [CGCM3.1(T47)]. Models that perform well over all North America also perform well over its western subregion. However, the model ranking is sensitive to the choice of climate variable. For another evaluation measure the method of self-organizing maps was applied to classify the characteristic daily patterns of sea level pressure over the region. The evaluation consists of correlating the frequencies of these patterns, as generated in GCMs, with the frequencies in the NARR over the baseline period. Most of the models are successful in simulating the frequencies of daily anomaly patterns from the 20-yr-average daily pattern. However, very few GCMs are able to reproduce the occurrences of characteristic daily weather patterns in the NARR on seasonal basis over the baseline period. In terms of relative performance, the three top performing models are the Meteorological Research Institute (MRI) CGCM2.3.2, ECHAM5–MPI-OM, and the Model for Interdisciplinary Research on Climate 3.2, high-resolution version [MIROC3.2(hires)]. The model skill in simulating daily synoptic patterns is not strongly linked to the skill in simulating the climatologies of selected variables. Despite the large scatter of model performance across all the metrics, some models consistently rank high [e.g., ECHAM5–MPI-OM and MIROC3.2(medres)]. Likewise, some models consistently rank low [e.g., the Community Climate System Model, version 3 (CCSM3) and the Goddard Institute for Space Studies Model E-R (GISS-ER)] independently of the evaluation measures, domain size, and climate variable of interest.

1. Introduction

Global climate models (GCMs) are commonly used tools for projecting future climate. GCM data contributing to the Fourth Assessment Report (AR4) of the Intergovernmental Panel on Climate Change (IPCC) have been collected in the Coupled Model Intercomparison Project Phase 3 (CMIP3) and are being used for impact studies of climate change on regional and local scales (e.g., Coquard et al. 2004; Brekke et al. 2008). Such studies share the problem of deciding which GCMs to use for further downscaling over a region of interest. A common way to address this problem is to evaluate model output against the reference data and then prequalify the models based on their ability to simulate climate in the region or variable of interest (e.g., Dettinger 2005; Milly et al. 2005; Tebaldi et al. 2005; Wang and Overland 2009; Barnett et al. 2008). Lacking reference data for the future, the climate model performance is evaluated against the present-day climate. Models that best simulate the present-day climate are assumed to yield the most credible projections of future climate, but there is no widely accepted set of measures for evaluating climate model performance (Gleckler et al. 2008). Most evaluation tools are statistical measures (e.g., mean error, root-mean square error, correlation, and

Corresponding author address: Valentina Radić, Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, BC V6T 1Z4, Canada.
E-mail: vradic@eos.ubc.ca

DOI: 10.1175/JCLI-D-11-00011.1

variance) for quantifying the differences between modeled and observed climatologies of a variable and region of interest. Gleckler et al. (2008) used statistical measures to evaluate the performance of 22 GCMs from CMIP3 over global and subglobal domains, and explored the possibility of establishing a single evaluation measure for overall model skill. Their results demonstrated that the skill of a model is highly sensitive to the choice of the statistical measures, climate variables, and spatial domains over which the evaluation is performed. These findings have discouraged the search for a single measure of overall model performance and encouraged the use of a wide range of evaluation measures.

Studies that evaluated the GCM simulation of mean climate in the region of interest (e.g., Gleckler et al. 2008; Reichler and Kim 2008; Pierce et al. 2009) confirmed that the multimodel ensemble average is superior to any individual model. Because the errors tend to be distributed around zero, averaging across models reduces the error. Similarly, if the measures for climate variability are averaged across the models, the mean ratio of model variance to observed variance approaches unity. When climate variability is considered in this way, the multimodel ensemble outperforms any individual model (Pierce et al. 2009). However, reduction of model errors by averaging across the ensemble does not increase confidence in the ability of the models to reproduce the climate features of interest (e.g., Tebaldi and Knutti 2007). Confidence in model performance is of particular importance to impact studies that commonly involve the additional steps of downscaling and process modeling.

In this study we evaluate the performance of 22 GCMs from CMIP3 over all North America and its western subregion using several different evaluation metrics. We direct our evaluation to the set of climate variables consisting of upper-air temperature and specific humidity, precipitation, geopotential heights, and mean sea level pressure. For these variables we adopt the evaluation measures from recent studies (e.g., Gleckler et al. 2008; Pincus et al. 2008; Walsh et al. 2008; Pierce et al. 2009) in order to quantify the biases in the modeled climatologies, specifically, the seasonal cycle and interannual variability. Although evaluation of modeled climatologies provides a reasonably comprehensive picture of model performance, it excludes some aspects of climate occurring on daily scales, such as the frequency and intensity of storms. The frequency of storms depends on the regional cyclonic activity, which is usually analyzed by looking at the patterns of sea level pressure. Therefore, in addition to analysis of modeled seasonal cycle and interannual variability, we evaluate how well the GCMs simulate the frequency of the daily synoptic patterns of sea level pressure in the region. We accomplish this by using a clustering algorithm known as self-organizing maps (SOMs) to identify and classify the characteristic synoptic patterns. SOMs are shown to be a powerful tool for model evaluation, allowing a detailed examination of the differences between simulated and observed atmospheric circulation (e.g., Finnis et al. 2008; Schuenemann and Cassano 2009).

Although the results of GCM evaluation over North America are relevant to a variety of climate change impact studies, our particular motivation is the impact on glaciers. Our overall goal is to select those GCMs that are best suited for modeling future volume changes of glaciers in southwestern Canada. Scattered over the Coast Mountains and Rocky Mountains, glaciers of this region cover $\sim$26 000 km^2, roughly 8 times the total area of glaciers in the European Alps. Glaciers in southwestern Canada are a significant source of water for agricultural, domestic, and industrial uses and for hydropower generation. Thus, the credibility of GCM climate projections is an important issue for hydrologic, energy and economic planning for the region, as well as for glacier modeling. The climate variables for our evaluation of GCM performance are all shown to be linked to glacier mass balance variability. For example, studies on glacier mass balance in this region have demonstrated that the observed winter mass balance strongly correlates positively with winter precipitation, whereas summer mass balance correlates negatively with summer air temperatures (e.g., Tangborn 1980; Hodge et al. 1998). Several studies have also demonstrated that variations in glacier mass balance are linked to variations in the atmospheric circulation, represented by regional patterns of sea level pressure, winds, and/or geopotential heights (e.g., Yarnal 1984; McCabe and Fountain 1995; Bitz and Battisti 1999; Shea and Marshall 2007; Arendt et al. 2009). Finally, observed accumulation and ablation for several glaciers within this region have been successfully modeled using upper-air climate variables, such as geopotential heights and specific humidity on 850 hPa and upper levels (e.g., Rasmussen and Conway 2001, 2003; Matulla et al. 2008).

Sections 2 and 3 summarize the data and methods used in this analysis, including a description of the SOM algorithm. Statistical measures are used to compare modeled versus observed seasonal cycle and interannual variability of selected climate variables; the SOM-based cluster analysis compares modeled versus observed frequencies of the synoptic patterns. In section 4 we present and discuss the results of the evaluation and rank the performance of GCMs according to the set of evaluation metrics. We further investigate relationships among model rankings that result from different metrics. We also analyze the sensitivity of model ranking to the size of the spatial domain and the choice of a baseline period over which the

TABLE 1. Model identification, originating center, and atmospheric resolution.

Model	Center and location	Atmosphere resolution
BCCR-BCM2.0	Bjerknes Centre for Climate Research (Norway)	T63 L31
CGCM3.1(T47)	Canadian Centre for Climate Modeling and Analysis (Canada)	T47 L31
CGCM3.1(T63)		T63 L31
CNRM CM3	Météo-France, Centre National de Recherches Meteorologiques (France)	T42 L45
CSIRO Mk3.0	Atmospheric Research (Australia)	T63 L18
GFDL CM2.0	U.S. Dept. of Commerce, NOAA	N45 L24
GFDL CM2.1	Geophysical Fluid Dynamics Laboratory (United States)	N45 L24
GISS-AOM	NASA Goddard Institute for Space Studies (United States)	90 × 60 L12
GISS-EH		72 × 46 L17
GISS-ER		72 × 46 L26
FGOALS-g1.0	LASG/Institute of Atmospheric Physics (China)	128 × 60 L26
INM-CM3.0	Institute for Numerical Mathematics (Russia)	72 × 45 L21
IPSL CM4	Institut Pierre Simon Laplace (France)	96 × 72 L19
MIROC3.2(hires)	Center for Climate System Research (The University of Tokyo)	T106 L56
MIROC3.2(medres)	National Institute for Environmental Studies, and Frontier Research Center for Global Change (JAMSTEC, Japan)	T42 L20
ECHO-G	Meteorological Institute of the University of Bonn, Germany Meteorological Research Institute of KMA, and Model and Data group (Germany and Korea)	T30 L19
ECHAM5–MPI-OM	Max Planck Institute for Meteorology (Germany)	T63 L32
MRI-CGCM2.3.2	Meteorological Research Institute (Japan)	T42 L30
CCSM3	National Center for Atmospheric Research (United States)	T85 L26
PCM		T42 L18
HadCM3	Hadley Centre for Climate Prediction and Research, Met Office (United Kingdom)	96 × 72 L19
HadGEM1		N96 L38

skill is assessed. Section 5 summarizes the results and provides a list of top-performing GCMs.

2. Model output and validation data

Our evaluation is based on the twentieth-century simulations by 22 GCMs from CMIP3 (Table 1), which have been archived at the Program for Climate Model Diagnosis and Intercomparison at Lawrence Livermore National Laboratory (LLNL). We used the same set of models as in the evaluation study by Gleckler et al. (2008). For most of these models the twentieth-century simulation is run from the 1800s with prescribed greenhouse gas concentrations and, in some cases, estimated sulfate aerosols and variable solar forcing (Randall et al. 2007, their Table 8.1). To make our evaluation analysis consistent, we use only one run for the GCMs that have multiple runs for a given scenario. We evaluate model performance for the climate variables that have been downscaled for glacier mass balance modeling (Jarosch et al. 2010) or linked to glacier mass balance via regression models (Matulla et al. 2009). These variables are monthly grids of precipitation (PR), geopotential heights at 500 hPa (Z500) and 850 hPa (Z850), specific humidity at 850 hPa (SH850), air temperature at 850 hPa (T850), and monthly and daily grids of sea level pressure (SLP).

The evaluation is performed by comparing GCM historical simulations with the North American Regional Reanalysis (NARR), which we take as the reference dataset. The NARR was generated by the National Centers for Environmental Prediction (NCEP) using surface, radiosonde, and satellite observational data assimilated into the Eta forecasting model. Output from more than 200 variables is available on an approximately 32-km grid, at 29 pressure levels and at 3-hourly intervals from the period 1979 to the present. A detailed description of the NARR is provided in Mesinger et al. (2006). We take the NARR as the reference data for GCM evaluation because, in a separate contribution (Jarosch et al. 2010), NARR air temperature and precipitation have been successfully validated against observations and downscaled for the mass balance modeling of glaciers in southwestern Canada. We assume that the NARR is a reasonable representation of the observed record for our intended impact study.

GCM historical simulations are compared with the NARR record for the period 1980–99. Although this 20-yr period is relatively short, it has the most complete and accurate observational data, largely because of the expansion of and advances in space-based remote sensing (Gleckler et al. 2008). This period has also been used in some of the recent studies on GCM evaluation (e.g., Gleckler et al. 2008; Pincus et al. 2008). To choose a 20-yr period from GCM historical simulations one needs to take into account that the individual years from the historical simulations are not expected to line up with those in the

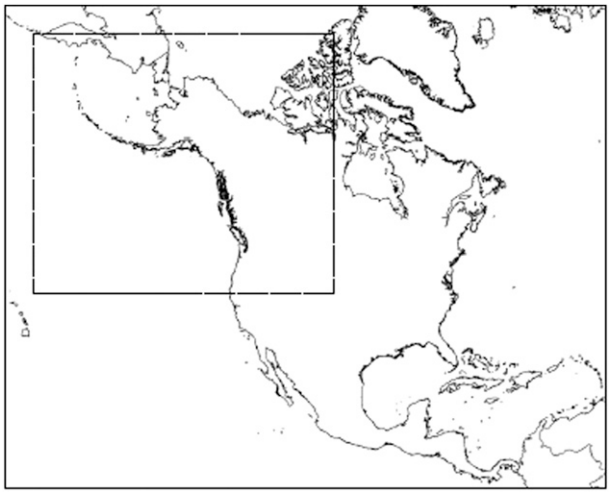

FIG. 1. Analysis domains: large domain (large rectangular) equivalent to the NARR domain and small domain (rectangle inside the large one).

observational record. This is because the only relationship between the GCM year and the real year is the equivalent amount of forcing from the natural and anthropogenic factors. To quantify how sensitive our analysis is to the choice of evaluation period we extract GCM data for two 20-yr time windows: 1980–99, which overlaps with the NARR period, and 1970–89, which is shifted 10-yr backward from the NARR period. To facilitate GCM intercomparison and validation against the NARR, all climate fields are interpolated to a common 10 × NARR grid (approximately 320 km × 320 km) in a conical conformal Lambert map projection.

3. Validation methods

a. Statistical metrics

Statistical metrics are used here to evaluate GCM simulations of the mean annual cycle and of interannual variability. We perform this analysis on two spatial domains: a "large" domain, equivalent to the original domain of the NARR, which covers all North America, and a "small" domain that occupies the northwest corner of the large domain (Fig. 1). For these two domains and our six selected climate variables we calculate relative model errors and variance ratios over the evaluation period as described below.

Following Gleckler et al. (2008) we define the relative model error using the root-mean-square difference between a simulated field F (GCM output) and a corresponding reference field R (NARR data). The root-mean square error (RMSE) is calculated as

$$\mathrm{RMSE}^2 = \frac{1}{W}\sum_i \sum_j \sum_t w_{i,j,t}(F_{i,j,t} - R_{i,j,t})^2, \quad (1)$$

where the indices i, j, and t correspond to the longitude, latitude, and time dimensions; and W is the sum of the weights ($w_{i,j,t}$), which for the spatial dimensions are proportional to gridcell area and for time are proportional to the length of each month. The gridcell area in our interpolated common grid is constant. The sums are accumulated over 12 months and separately over each of the two spatial domains. For each climate field Gleckler et al. (2008) defined a typical model error as the median of all RMSE calculations over all the GCMs. Thus, in our ensemble of 22 GCMs there are 22 RMSE calculations for one climate variable. Relative model performance, for a given model and climate variable, is then defined as a difference between the RMSE and the typical model error, normalized by the typical model error. Normalizing the RMSE calculations in this way yields a measure of how well a given model compares with the typical model. For example, if the relative error has a value of 0.5 then the model RMSE is 50% larger than the typical model error.

In an attempt to define an optimal overall index of model performance in simulating mean annual cycle climatology, Gleckler et al. (2008) introduced a model climate performance index (MCPI). This was done by averaging relative errors for each model across all climate variables considered. In our work, the MCPI is calculated by averaging the relative errors over our six selected climate variables.

Another statistical measure for evaluating GCM performance examines how well the model simulates the interannual variability of the climate variables of interest. This is analyzed by the variances of monthly mean anomalies, computed relative to the monthly climatology for the 20-yr baseline period. The ratio of simulated to observed variances for the two domains (small and large) is calculated for each GCM and climate variable. A variance ratio (GCM vs NARR) close to unity indicates that the variance of simulated monthly anomalies, for a given climate variable, compares well with NARR, whereas a lower ratio suggests too little simulated variability and a higher ratio implies too much.

Continuing to follow Gleckler et al. (2008), we calculate the model variability index (MVI), which serves as an overall index of model performance for simulating the interannual variability. MVI is defined as

$$\mathrm{MVI} = \sum_{n=1}^{N}\left(\beta_n - \frac{1}{\beta_n}\right)^2, \quad (2)$$

where β^2 is the ratio of simulated to observed variance and N is the total number of climate variables (in the

present work $N = 6$). Defined this way, the MVI is positive, with smaller values indicating better overall agreement between modeled and reference data.

b. Self-organizing maps technique

Here we follow the methodology from Finnis et al. (2008) and Schuenemann and Cassano (2009) and compare daily SLP patterns in 22 GCMs to those in the NARR. The patterns are extracted by the method of SOMs. We aim to find the GCMs that best reproduce the occurrences of NARR synoptic-scale patterns over our large and small domains.

SOMs are a common type of unsupervised artificial neural network particularly adept at pattern recognition and classification, and in many respects are analogous to more traditional forms of cluster analysis. The main difference from the other forms of cluster algorithms is that in SOMs no assumptions regarding the resulting patterns are made by the user. Kohonen (2000) offers an explanation of the development and details of the SOM algorithm. The method is used in a wide range of disciplines (Kaski et al. 1998; Oja et al. 2002), but here we use it to classify patterns in the climate data, an approach that has been successfully demonstrated in previous studies (e.g., Hewitson and Crane 1994, 2002; Malmgren and Winter 1999; Cavazos 2000; Ambroise et al. 2000; Hsu et al. 2002; Hong et al. 2004, 2005; Finnis et al. 2008). The SOM method offers several advantages over principal component analysis, for example in revealing real synoptic patterns and in feature extraction (Reusch et al. 2005; Liu et al. 2006).

Briefly the SOM algorithm proceeds as follows: the input data consists of climate fields each of which is converted to a row vector. The output is an SOM that consists of a prescribed number of nodes that represent the archetypal patterns in the input data. Initially, the SOM consists of random nodes, where each node has an assigned weight vector and a position in the 2D map space. The procedure for placing a vector onto the map is to find the node with the closest weight vector to the input vector and to assign the map coordinates of this node to the vector. The node with the closest weight vector (in Euclidian space) is called the ''best matching unit.'' The next step is to update the nodes in the neighborhood of the best matching unit by pulling them closer to the input vector, for example:

$$\mathbf{W}(t+1) = \mathbf{W}(t) + \Theta(t)\alpha(t)\,[\mathbf{V}(t) - \mathbf{W}(t)], \quad (3)$$

where t is the current iteration, $\mathbf{W}$ is the weight vector, $\mathbf{V}$ is the input vector, Θ is restraint due to distance from the best matching unit (usually called the neighborhood function), and α is a time-dependent learning restraint. Repeating this process for all the input data is referred to as training the SOM. Training is performed for a chosen iteration limit. After the training process, individual SOM nodes represent characteristic patterns in the original data. The amount of original information retained depends primarily on the size of the SOM (i.e., the number of nodes), with smaller sizes producing broad generalizations of the input dataset, and larger sizes capturing increasingly fine details. The essential feature of the SOM is that neighboring nodes represent similar patterns, while those that are placed farther apart are more dissimilar.

For the period 1980–99 we apply the SOMs technique to daily SLP anomalies from the NARR and 21 GCMs [note that daily data from the Hadley Centre Global Environmental Model version 1 (HadGEM1) model is not available in the LLNL archive]. Prior to the SOM analysis, daily SLP are interpolated to the common grid (320 km $\times$ 320 km). The input data for the SOM training process are temporal SLP anomalies calculated by subtracting the daily averaged SLP over the 20-yr baseline period from the daily SLP at each grid point, for the NARR and each GCM separately. As an alternative input for SOM analysis, we also calculate spatial SLP anomalies by subtracting the domain-averaged daily SLP from the SLP at each grid point. The spatial SLP anomalies have been more commonly used in previous SLP pattern analysis (e.g., Finnis et al. 2008; Schuenemann and Cassano 2009), allowing the SOM algorithm to focus on the SLP gradients rather than the varying magnitudes of SLP from day to day.

Our SOM analysis uses the Matlab SOM Toolbox (Vesanto et al. 2000) and assigns tunable parameters of the SOM training process (e.g., neighborhood function and radius, type of training, initialization of weight vectors, number of iterations, etc.). To ensure the robustness of our analysis we train SOMs using several different map sizes (number of patterns) with varying parameters. Following the guidelines from Liu et al. (2006) we accept a set of parameters that minimizes the average quantization error (i.e., the average distance between each data vector and the best matching unit) and topographic error (i.e., the percentage of the data vectors for which the first and the second best-matching unit are not the neighboring nodes). Our set of optimized parameters in Matlab SOM Toolbox consists of the following: hexagonal lattice, sheet SOM shape, linearly initialized weights, bubble neighborhood function with initial and final neighborhood radii of 3 and 1, and batch training performed over 5000 iterations. With this set of parameters we train the SOMs on a seasonal basis [winter months of December–January–February (DJF), spring months of March–April–May (MAM), summer months of June–July–August (JJA), and autumn months of September–October–November (SON)]. Prior

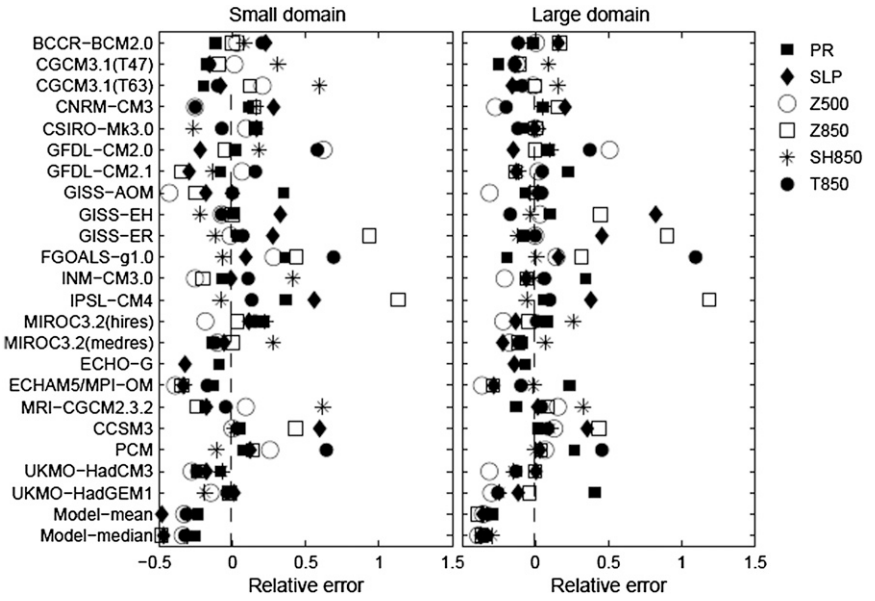

FIG. 2. Relative errors over (left) the small and (right) the large domain for the six climate variables: PR, SLP, Z500, and Z850, specific humidity at 850 hPa (SH850) and air temperature at 850 hPa (T850).

to the SOM training, the SLP anomalies are converted into row vectors (from a 2D spatial field into a 1D array). Hence, our input data to the SOM training consists of row vectors from the NARR followed by the row vectors from one of the 21 GCMs, making in total 21 sets of input data. The training is performed independently for each set and each season. The SOMs for the temporal SLP anomalies are produced for the large and small spatial domain, while the SOMs for the spatial SLP anomalies are produced only for the small domain.

For a given size of SOM one can use a different set of optimized parameters. We found that, once the SOM map size is set, the sensitivity of pattern recognition in the SOM training to the choice of optimized parameters is small. However, our evaluation of GCM performance might be sensitive to the choice of the SOM size. We experimented with different SOM sizes in order to find a reasonable compromise between detail and interpretability of the SLP patterns characteristic for each season. Our final choice for both spatial domains is to use three SOM sizes: 4×3, 4×4, and 5×4. Having more than one SOM size allows us to test the sensitivity of model evaluation to the size of SOM.

4. Results

a. Simulation of mean annual cycle

In Fig. 2 we provide a summary of relative errors over the small and large domain, for our six selected climate variables (PR, SLP, Z500, Z850, SH850 and T850). The results are shown for the monthly climatology computed over 1980–99 from each GCM. Models with negative relative errors are in better agreement with the reference data (NARR) than the typical model. Thus, the more negative the relative error, the better the skill of the model in simulating the mean annual cycle of the selected variable. Note that the ECHAM and the global Hamburg Ocean Primitive Equation (ECHO-G) model only have data available for two of the climate variables (PR and SLP) from our selected set. As illustrated in Fig. 2, some models perform better than others, although no model scores above average or below average for all the climate variables. For some models the range of relative errors for a set of climate variables is narrower than for other models. For example, models with a narrow range over the small domain are the HadGEM1, the third climate configuration of the Met Office Unified Model (HadCM3), and the ECHAM5–Max Planck Institute Ocean Model (MPI-OM), whereas L'Institut Pierre-Simon Laplace Coupled Model, version 4 (IPSL CM4), Goddard Institute for Space Studies Model E-R (GISS-ER), and the Geophysical Fluid Dynamics Laboratory Climate Model version 2.0 (GFDL CM2.0) have a wide range. Several models have more relative errors greater than 0.5 over the small domain than is the case over the large domain. This suggests that simulating mean annual cycle over the small domain is more challenging than over the large domain. However, the correlation between the relative errors over large and small domains is high ($r > 0.7$) for all the climate

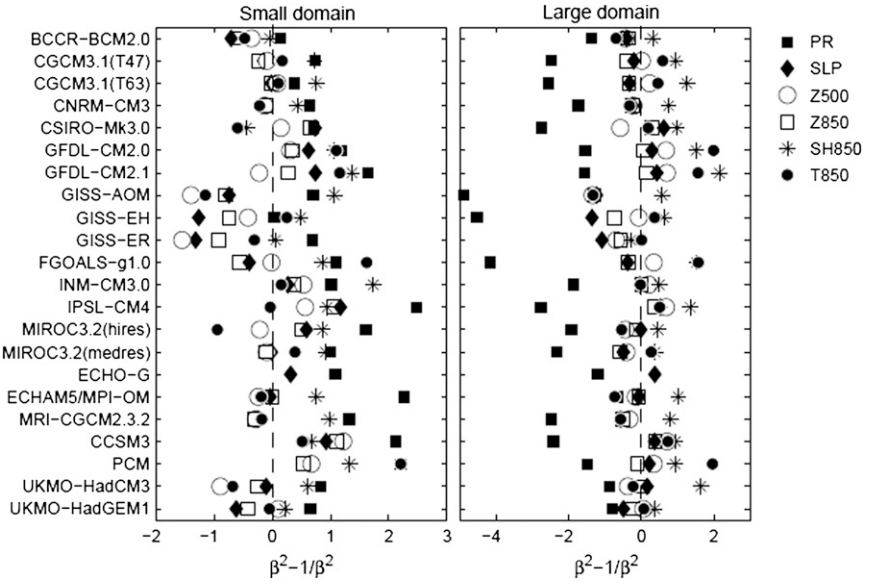

FIG. 3. Values of $\beta^2 - 1/\beta^2$, where β^2 represents the ratios of simulated (GCM) to reference (NARR) variances, over the small and large domain, for the six climate variables in Fig. 1.

variables except precipitation. Thus, the relative model performance in simulating the mean annual cycle of the selected five variables over the small domain is comparable to that for the large domain.

The results presented in Fig. 2 are almost identical when using two different baseline periods (1970–89 and 1980–99) over which the mean annual cycle in GCMs is calculated. Therefore, the model performance in simulating the mean annual cycle of our 6 variables is not sensitive to the 10-yr shift in the baseline period. For both domains and all climate variables, the multimodel mean and median have relative errors that are smaller than the typical model error. Furthermore, in most cases the multimodel mean and median score better than any individual model, a result that has been previously reported (e.g., Taylor and Gleckler 2002; Gleckler et al. 2008). An exception here is the performance of GISS Atmosphere–Ocean Model (GISS-AOM) and ECHAM5–MPI-OM, which score better than the multimodel mean and median for simulating geopotential height at 500 hPa (Z500) over the small domain. ECHAM5–MPI-OM also scores best over the small domain when simulating specific humidity at 850 hPa (SH850).

Looking at the relative errors of each individual climate variable across all the models, we note that some climate variables have a distribution with severe outliers. For example, geopotential height Z500 for IPSL CM4 model over both domains has the largest relative errors in the set (the values are between 2 and 3, beyond the x-axis scale illustrated in the figure). Both geopotential heights, Z850 and Z500, have larger scatter of relative errors across the models, than the other climate variables. The large scatter of relative errors for Z850 is mainly caused by the large errors from IPSL CM4 and GISS-ER.

We also examine whether there is a relationship between the performances across climate variables for each model, where a strong relationship would indicate redundancy in the evaluation. Correlating the relative errors from a pair of two different variables, we find significantly positive correlation (at the 95% confidence level) for the pairs (SLP, Z850) and (Z500, Z850) over both domains. For the small domain there is also a significant relationship between how well the model simulates PR and SLP ($r = 0.56$), and PR and Z850 ($r = 0.60$). In contrast, the relationship between how well the model simulates SLP and Z500 is weaker ($r = 0.46$), a result found at the global scale by Gleckler et al. (2008).

b. *Simulation of interannual variability*

Here we analyze simulated interannual variability by examining variances of monthly mean anomalies that are computed relative to monthly climatology. The results are shown for the monthly climatology computed over 1980–99 from each GCM. Figure 3 illustrates the ratio of simulated (GCM) to reference (NARR) variances, β^2 for each climate variable in the two domains. Values are expressed as $\beta^2 - 1/\beta^2$ so that the values closer to zero indicate that the variance of simulated monthly anomalies compares well with the NARR variance. Negative values reveal low simulated variability, while positive values correspond to high simulated variability.

Similar to the results in the previous section, some models score better than the others, while no model is superior for all the climate variables. Across all models the correlation of the variance ratios between the large and small domains is high ($r > 0.75$) for every climate variable except PR ($r = 0.28$) and SH850 ($r = 0.35$).

Evaluation of the multimodel mean and median is omitted from this analysis because averaging over the models reduces the variability and leads to an unrealistically small variance. Averaging the $\beta^2 - 1/\beta^2$ values in Fig. 3 across all the models for each climate variable reveals that the variances of SLP, Z850, and Z500 are better simulated than the variances of the remaining variables. All models show low variance ratios for precipitation over the large domain (average value for $\beta^2 - 1/\beta^2$ across all models is -2.46), suggesting too little simulated variability. The precipitation over the small domain shows too high simulated variability (average value for $\beta^2 - 1/\beta^2$ is 1.41). This reveals an inconsistency in model skill when simulating the variance of monthly precipitation over the two domains.

Correlating the variance ratios across the models for different pairs of variables we find significantly positive correlation (at the 95% confidence level) occurring over both domains for the following pairs: (PR, SLP), (SLP, Z500), (SLP, Z850), (Z500, Z850), and (PR, Z850). Especially high positive correlation coefficient ($r > 0.9$) is between SLP and Z850. Neither domains show significant correlation between PR and T850 ($r < 0.35$). Additionally, SLP and SH850 have significantly positive correlations over the large domain ($r = 0.60$), while only weak correlations over the small domain ($r = 0.36$).

We repeat this analysis using the monthly climatology computed over 1970–89 for each GCM. The correlation between the $\beta^2 - 1/\beta^2$ values in Fig. 3 over the small domain and the values calculated over 1970–89 period is very high ($r > 0.90$) for each climate variable except for SH850 ($r = 0.78$). Nevertheless, the results over large domain show larger disparity. Here the correlation between values $\beta^2 - 1/\beta^2$ for SH850, derived from different baseline periods, drops to 0.57, while for Z500 $r = 0.74$, for T850 $r = 0.80$ and for the remaining variables $r > 0.90$. Thus, for all variables except SH850 the model performance, in term of simulating the interannual variability, has a small sensitivity to the 10-yr shift in the baseline period.

c. Simulation of occurrences of synoptic patterns

Figures 4a–f illustrates characteristic daily SLP anomaly patterns for winter and summer over the small and large domain for the period 1980–99. The patterns are derived from the SOM training with map size 4×4 using temporal (Figs. 4a,b,e,f) and spatial (Figs. 4c,d) SLP anomalies from the NARR and one GCM (in this example CCSM3). The SOM of the temporal SLP anomalies shows the 2D distribution of SLP daily anomalies from the 20-yr average of daily SLP patterns (Figs. 4a,b,e,f). During winter months the most common daily pressure system is the Aleutian low. As a result, the majority of SOM patterns in winter represent characteristic anomalies in the strength of the Aleutian low (Fig. 4a). During summer months the Pacific tends to be dominated more by the subtropical anticyclone. However, as it was the case for winter months, the majority of the 2D anomaly patterns in summer also reveal positive and negative biases in the strength of the Aleutian low (Fig. 4b). Nevertheless, the amplitude of these anomalies in summer is smaller than in winter. SOM patterns for spring and autumn months (not shown in the figure) depict similar patterns, again showing the dominance of the Aleutian low in the temporal SLP anomaly patterns.

While SOM of the temporal SLP anomalies reveal the daily anomalies from the 20-yr-average pattern of daily SLP, more information on actual weather maps is given in the SOM of the spatial SLP anomalies. For example, the SOM nodes in Fig. 4d depict the dominant weather patterns in winter and enable us to follow the development of winter cyclones over the map. Most patterns are characterized by low pressure centers in the North Pacific in the vicinity of the climatological Aleutian low (e.g., nodes [4, 1], [4, 2], [4, 3], and [4, 4]). Some nodes illustrate the cyclones that are shifted southward as anticyclones become more present over the Beaufort Sea (e.g., nodes [3, 1] and [3, 2]), and some nodes show the weaker low pressure systems that are shifted into the Gulf of Alaska (e.g., [2, 2] and [2, 3]) or shifted into the interior of Alaska (e.g., [2, 4] and [1, 4]). In the SOM for summer (Fig. 4d), the circulation types are mainly dominated by the northeastern Pacific subtropical anticyclone, and the pressure gradient over the domain is generally smaller than in winter. SOMs for the transitional seasons (MAM and SON), which are not presented in the figure, produce a range of patterns characteristic of winter (e.g., North Pacific cyclones) and summer (e.g., high pressure centers in the North Pacific that are located slightly southward from their positions in summer).

Having created the SOMs of characteristic SLP anomaly patterns for the NARR and each GCM, the next step is to evaluate model performance. Good models would recreate the same synoptic patterns that take place in the real atmosphere, here represented by the NARR, at the same frequencies of occurrence. We calculate the node frequency (in %) as the total number of days with a certain pattern (node) divided by the total number of days for that particular season over the 20-yr baseline period. The success of the model depends on how well these frequencies from GCM simulations correlate with frequencies from the NARR.

In Fig. 5 we plot the frequency of each node from the NARR and a GCM (here CCSM3). The node coordinates

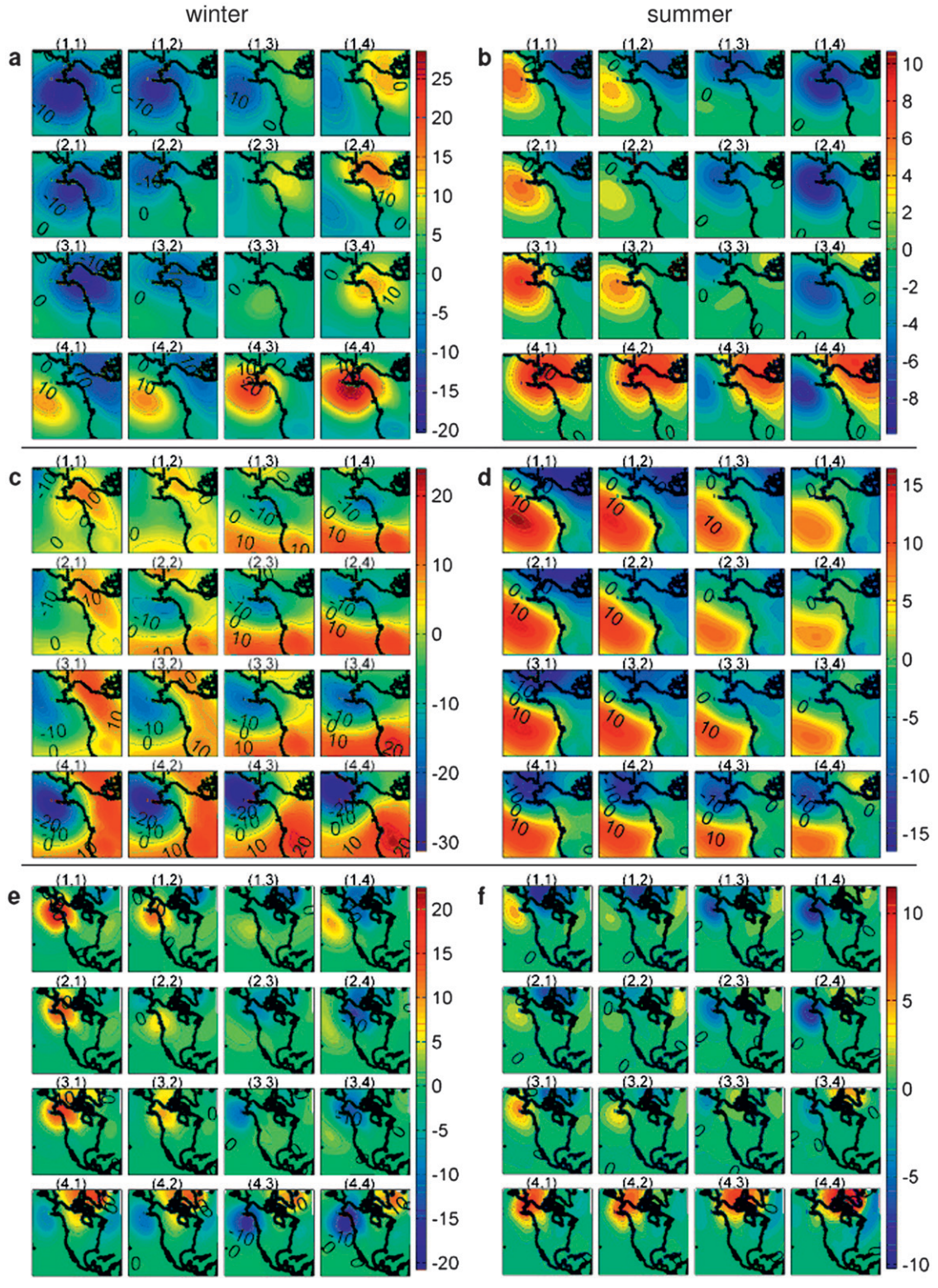

FIG. 4. The 4 × 4 SOMs of SLP anomalies (hPa) trained from the NARR and one GCM (CCSM3) over the baseline period 1980–99. Patterns of temporal SLP anomalies over the small domain (a) in winter (DJF) and (b) in summer (JJA). Patterns of spatial SLP anomalies over the small domain (c) in winter and (d) in summer. Patterns of temporal SLP anomalies over the large domain (e) in winter and (f) in summer.

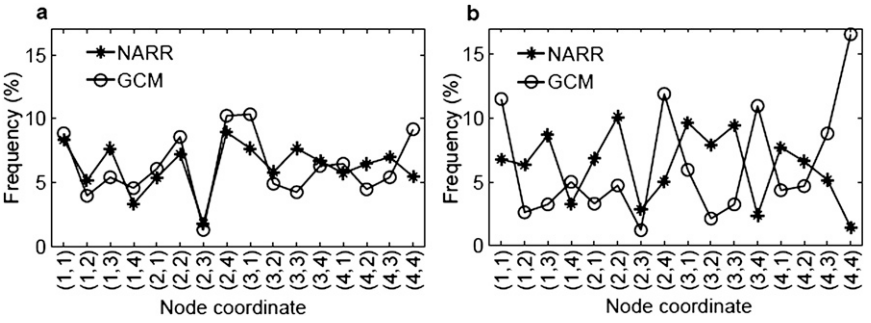

FIG. 5. Comparison of model performance for temporal and spatial SLP anomalies. NARR and GCM (CCSM) frequency of occurrences (%) for each node on the 4 × 4 SOM of winter SLP anomalies over the small domain. (a) Node coordinates correspond to the node coordinates of SOM in Fig. 4a. (b) Node coordinates are from SOM in Fig. 4c.

in Figs. 5a,b, correspond to the node coordinates in Figs. 4a,c, respectively. For the patterns of temporal SLP anomalies (Fig. 5a), a series of node frequencies in the GCM has significant positive correlation (at the 95% confidence level) with the series of node frequencies in the NARR ($r = 0.68$). Very positive correlation between these series means that each pattern (showing a characteristic 2D distribution of biases from the 20-yr average of daily SLP) in the GCM occurs as often as the equivalent pattern occurs in the NARR for a given season over the baseline period. For the patterns of spatial SLP anomalies (Fig. 5b) the correlation is significant but negative ($r = -0.51$). Very negative correlation here means that the occurrence of SLP spatial patterns in the GCM is almost in antiphase with the occurrence of the equivalent patterns in the NARR. In other words, the spatial SLP patterns that happen very frequently in the NARR happen only occasionally in the GCM and vice versa. We perform this correlation analysis between the NARR and each GCM, for each season (DJF, MAM, JJA, SON), each SOM size (4 × 3, 4 × 4, 5 × 4), and for both spatial domains. The results (r values) for the temporal and spatial 4 × 4 SOMs over the small domain and for all seasons are shown in Table 2. Correlations that are significantly greater than zero (at the 95% confidence level) are in boldface. As illustrated in the table, correlation varies widely from one model and season to another, ranging from near-perfect positive correlation to negative correlation. Analysis for other SOM sizes (4 × 3 and 5 × 4) gave comparable ranges for r values.

For temporal SLP anomaly patterns, 15 out of 21 GCMs have significantly positive correlations with the NARR frequencies across all the seasons and all SOM sizes (4 × 3, 4 × 4, 5 × 4). Specifically, for the small domain, winter has 95% of cases with significantly positive r values, spring has 92%, autumn has 82%, and summer has 73%. For the large domain, summer has 100% of cases with significantly positive correlations, followed by winter with 94% of cases, while spring and autumn both have 90%. This high number of significantly positive correlations means that almost every GCM is able to reproduce the frequencies of the temporal SLP anomaly patterns in the NARR on seasonal basis over the baseline period. Further analysis showed that the success in reproducing these frequencies depends on how well a GCM simulates the 20-yr average of seasonal SLP pattern. The larger the difference (in terms of RMSE) between the average seasonal SLP patterns in a GCM and the NARR, the larger the difference between the frequencies of characteristic anomaly patterns in the GCM and the NARR.

While the frequencies of temporal SLP anomaly patterns are well simulated across all GCMs, analysis for spatial SLP anomaly patterns shows a very small number of significantly positive correlations between GCM and the NARR frequencies. Across all the models and all SOM sizes, winter has 27% of cases with significantly positive correlation, followed by spring with 22%, autumn with 11%, while summer has only 2%. This relatively small number of significantly positive correlations means that a very few GCMs are able to reproduce the frequencies of spatial SLP anomaly patterns in the NARR for any season over the baseline period. Furthermore, frequencies of the patterns from some GCMs have significantly negative correlation with those from the NARR. This reveals not only the poor GCM performance in reproducing the frequency of spatial SLP patterns in the NARR, but shows that the occurrences of patterns in some GCMs are almost in antiphase with those in the NARR. For example, both the Commonwealth Scientific and Industrial Research Organisation Mark version 3.0 (CSIRO Mk3) model with $r = -0.47$ and IPSL CM4 with $r = -0.62$ overestimate the occurrence of a strong Aleutian low in the winter season, and therefore underestimate the occurrence of other characteristic patterns in the season. The summer season has the largest

TABLE 2. Correlation coefficients r between node frequencies of 4×4 SOM in the NARR and each GCM, on seasonal basis (DJF, MAM, JJA, and SON). SOMs are given for the temporal and spatial SLP anomalies over the small domain. Bold font marks the correlations significantly > 0 (at the 95% confidence level).

	Temporal SLP patterns				Spatial SLP patterns			
Model	DJF	MAM	JJA	SON	DJF	MAM	JJA	SON
BCCR-BCM2.0	**0.58**	**0.55**	**0.60**	**0.74**	0.10	0.34	−0.71	**0.53**
CGCM3.1(T47)	**0.64**	**0.56**	**0.83**	**0.60**	0.29	0.29	0.13	−0.03
CGCM3.1(T63)	**0.79**	**0.60**	**0.72**	**0.57**	0.42	0.15	0.22	−0.46
CNRM-CM3	**0.67**	**0.60**	0.38	**0.63**	**0.54**	−0.06	−0.67	0.44
CSIRO Mk3.0	**0.85**	**0.79**	**0.74**	0.50	−0.47	0.47	−0.09	0.00
GFDL CM2.0	**0.79**	**0.66**	**0.94**	0.43	**0.60**	−0.11	−0.33	0.19
GFDL CM2.1	**0.76**	**0.84**	**0.72**	0.46	**0.73**	0.49	−0.17	0.39
GISS-AOM	0.30	**0.58**	0.15	**0.80**	0.32	0.27	−0.68	0.43
GISS-EH	**0.59**	0.46	0.34	**0.65**	0.39	−0.43	−0.83	0.06
GISS-ER	**0.63**	0.26	0.17	**0.75**	−0.33	−0.57	−0.80	−0.43
FGOALS-g1.0	0.50	**0.75**	−0.04	**0.86**	−0.28	−0.17	−0.62	0.28
INM-CM3.0	**0.81**	**0.78**	**0.69**	**0.63**	−0.27	0.47	−0.55	−0.36
IPSL CM4	**0.72**	**0.73**	**0.72**	**0.61**	−0.60	−0.27	−0.73	−0.53
MIROC3.2(hires)	**0.87**	**0.90**	**0.83**	**0.78**	−0.10	**0.50**	0.13	0.47
MIROC3.2(medres)	**0.72**	**0.75**	**0.85**	**0.85**	−0.14	0.33	0.22	−0.23
ECHO-G	**0.77**	**0.76**	**0.85**	**0.81**	**0.53**	0.29	−0.45	**0.71**
ECHAM5–MPI-OM	**0.80**	**0.63**	**0.75**	**0.72**	**0.55**	0.34	0.42	0.12
MRI-CGCM2.3.2	**0.75**	**0.67**	**0.80**	**0.61**	**0.51**	**0.65**	−0.11	0.10
CCSM3	**0.68**	0.05	0.50	0.45	−0.51	−0.70	−0.46	−0.63
PCM	**0.83**	**0.84**	**0.77**	0.36	0.01	−0.40	−0.21	−0.51
HadCM3	**0.73**	**0.77**	**0.77**	**0.56**	−0.27	0.02	−0.62	0.12

number (49%) of cases with significantly negative r values, and most models in those cases underestimate the occurrence of a weak low pressure over the Bering Sea or Arctic North America combined with northeastern Pacific subtropical anticyclone. Some of these models {e.g., Flexible Global Ocean–Atmosphere–Land System Model gridpoint version 1.0 (FGOALS-g1.0) and Model for Interdisciplinary Research on Climate 3.2, high-resolution version [MIROC3.2(hires)]} produce a higher frequency of the summer pattern characterized by a high pressure center over Arctic North America than does the NARR (e.g., pattern similar to the node [4, 4] in Fig. 2d).

How robust is this correlation analysis to the choice of the baseline period over which the SOM patterns and their frequencies are derived? To answer this we repeat the SOM training using the SLP temporal and spatial anomalies from the period 1970–89 for each GCM, while the NARR period is kept the same (1980–99). We then compare the new r values between GCM and the NARR frequencies across all seasons and all SOM sizes. The results for spatial SLP patterns show that the correlation between the old and new r values is larger than 0.85, proving a strong link between the r values from the two baseline periods. However the results for temporal SLP anomaly patterns reveal some weaker links. For example, the correlation between the old and new r values over the small domain is the highest for summer (average correlation over the 3 SOM sizes is 0.87), followed by statistically significant correlations for spring (0.73) and autumn (0.63), while winter has statistically insignificant correlation of 0.38. Over the large domain, the correlation is also the highest for summer (0.63), followed by autumn (0.58) and spring (0.50), whereas winter again has statistically insignificant correlation of 0.06. We conclude that for all the seasons except winter, the model skill in simulating the frequencies of daily SLP anomaly patterns is not significantly sensitive to the 10-yr shift in the baseline period. A possible explanation for the relatively high sensitivity in winter is found by looking at the North Pacific index. This index depicts the fluctuations in the intensity of the Aleutian low over winter months, as well as a regime shift in these fluctuations as part of Pacific decadal time-scale variation (Trenberth and Hurrell 1994). Even if a GCM is able to simulate the interdecadal variability of the Aleutian low regime there is no reason to expect that the shift in the regime, as simulated in GCM, would line up with the regime shift in the NARR. Thus, a decade shift in the baseline period would make the largest impact on model evaluation in winter season, improving the performance of some GCMs that underperformed in the original baseline period, and/or worsening the performance of some GCMs which performed better in the original baseline period.

To rank the models based on the correlations between the node frequencies from the NARR and GCMs, we introduce a set of evaluation measures. Our first measure

is a correlation measure M_C defined as the mean correlation across all the seasons and all three SOM sizes:

$$M_C = \frac{1}{m}\sum_{i=1}^{m} r_i, \tag{4}$$

where m is the product of the seasons and all three SOM sizes ($m = 12$) and r is the correlation coefficient. The larger this measure, the closer the model performance is to the NARR performance.

To account for the total number of significant positive correlations we define a significance measure:

$$M_S = \sum_{i=1}^{m} \delta_i \begin{cases} \delta_i = 1 & \text{if} \quad r_i \geq r_0 \\ \delta_i = 0 & \text{if} \quad r_i < r_0 \end{cases}, \tag{5}$$

where r_0 is the threshold value for the correlation significantly larger than zero at the 95% confidence level (derived from a t test). Similarly to M_C, the larger this measure, the closer the model performance is to the NARR performance.

Finally we define a rank measure M_R following the approach in Schuenemann and Cassano (2009), by ranking the models from best to worst according to the correlation with the NARR frequencies, for each season and each SOM size separately. Thus, the top-performing model has rank = 1, while the bottom performing has rank = 21. These ranks are then totaled across all the seasons and all three SOM sizes. The smaller the sum, the better is the model performance. To facilitate comparison with the measures M_C and M_S we define this measure as

$$M_R = 1 - \frac{1}{252}\sum_{i=1}^{m} \text{rank}_i, \tag{6}$$

so that the larger the value, the closer the model performance is to the NARR performance. In Eq. (6) the total sum of ranks across all the seasons and all the SOM sizes ($m = 12$) is divided by the product between m and the bottom rank (rank = 21).

Figure 6 shows the three measures assessed over the small domain for temporal and spatial SLP anomaly patterns, and over the large domain for temporal SLP anomaly patterns. As illustrated in the figure, the evaluation of the model performance depends on the choice of the evaluation measure even when the same climate features are analyzed. Looking only at measure M_S one can see the discrepancy in model performance between simulating the temporal SLP anomaly patterns on one hand and spatial SLP anomalies patterns on the other. This discrepancy is also revealed with M_C measure, while the M_R gives only the information of the relative model

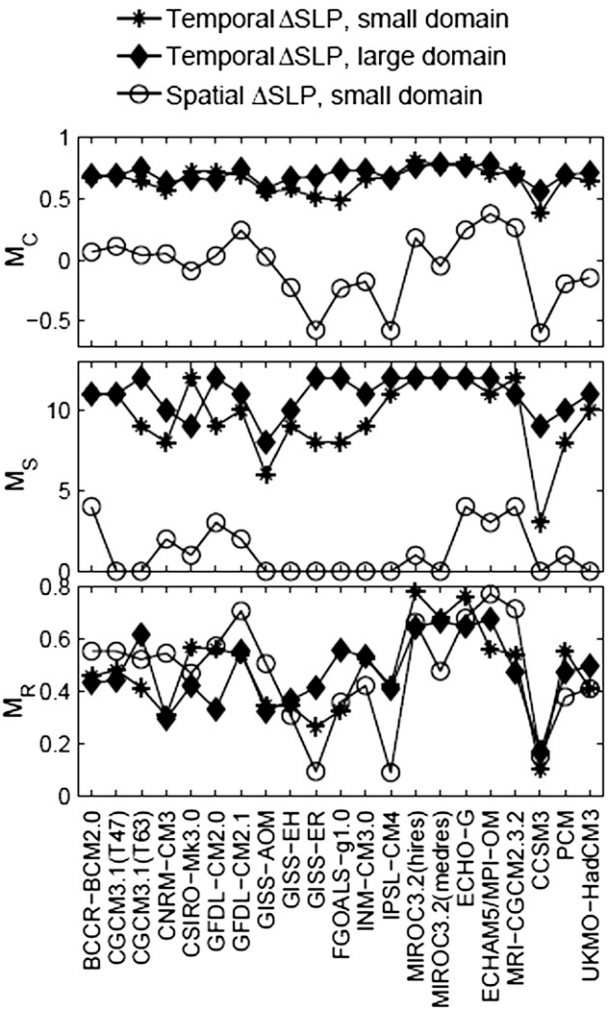

FIG. 6. (top to bottom) Correlation measure M_C, significance measure M_S, and rank measure M_R for each model. The measures are derived for three different cases: SOMs of temporal SLP anomalies over the small domain, SOMs of temporal SLP anomalies over the large domain, and SOMs of spatial SLP anomalies over the small domain.

performance. Looking specifically at the results for temporal SLP anomaly patterns over the small domain we find that the values across all three measures have significantly positive correlations at the 95% confidence level ($r > 0.80$). The largest r value of 0.96 is between the measures M_C and M_R. Similar r values across the three measures are also found for spatial SLP anomaly patterns over the small domain and temporal SLP anomaly patterns over the large domain. Thus, the sensitivity of the model ranking to the choice of the three measures is small.

The correlation between the M_S values for spatial and M_S values for temporal SLP anomaly patterns is not significant at the 95% confidence level ($r = 0.41$). This

means that the model skill in simulating the temporal SLP anomaly patterns is not significantly linked to the skill in simulating the spatial SLP patterns. However, the same correlation analysis applied on the other two measures (M_C and M_R) reveals that the correlations are significantly positive ($r > 0.64$). In terms of simulating temporal SLP anomaly patterns for the two different domains, there is a significant correlation between the M_S values for the small and M_S values for the large domain ($r = 0.52$). Both correlations are also significantly positive ($r > 0.63$) when measures M_C and M_R are used. Thus, the sensitivity of the model ranking to the choice of our two domains is small.

d. Model ranking across all measures

Here we rank the model performance using all the measures we described in previous sections. The ranking is sorted in ascending order where the top-performing model has a rank = 1. First, we rank the GCMs according to how well they simulate the mean annual cycle and interannual variability, using the MCPI and MVI, respectively (Fig. 7). For the small domain, the three top-performing models are ECHAM5–MPI-OM, HadCM3 and GFDL CM2.0 based on MCPI, and Centre National de Recherches Météorologiques (CNRM), CGCM3.1(T63), and HadGEM1 based on MVI. For the large domain, the three top-performing models are ECHAM5–MPI-OM, HadCM3 and CGCM3.1(T47) based on MCPI, and HadGEM1, ECHAM5–MPI-OM, and the Bjerknes Centre for Climate Research (BCCR) based on MVI.

We correlate the ranking according to MCPI with the ranking according to MVI to investigate whether there is any relationship between the model ability to simulate the mean climate and its ability to simulate the interannual variability. For the small domain $r = 0.28$ is not significantly different from zero at the 95% confidence level, while for the large domain $r = 0.54$ is just above the threshold to be significantly different from zero. Consistent with our results Gleckler et al. (2008) found at best a weak relationship between MCPI and MVI, derived from seven selected climate variables over different domains.

Next we rank the models according to their simulation of frequencies of daily synoptic patterns using the three measures (M_C, M_S, and M_R). In Fig. 8 we illustrate the model ranking according to the sum of the measures M_C and M_S over all seasons for three different SOM sizes (4×3, 4×4, 5×4). Depending on the selected SOM size the model rank can change by as much as 16 places (e.g., HadCM3 for the large domain), while the average change of rank per model is 4 places for the small domain and 5 places for the large domain. The positive correlation between the model ranking from three different SOM sizes is statistically significant ($r > 0.60$) for the small domain. The same correlation analysis over the large domain shows significantly positive correlations ($r > 0.54$) between all the SOM sizes except for the pair 4×3 and 5×4 where $r = 0.40$. We conclude that the sensitivity of the model ranking to the choice of the three SOM sizes is greater than ideal, but small enough for this evaluation of GCM skill.

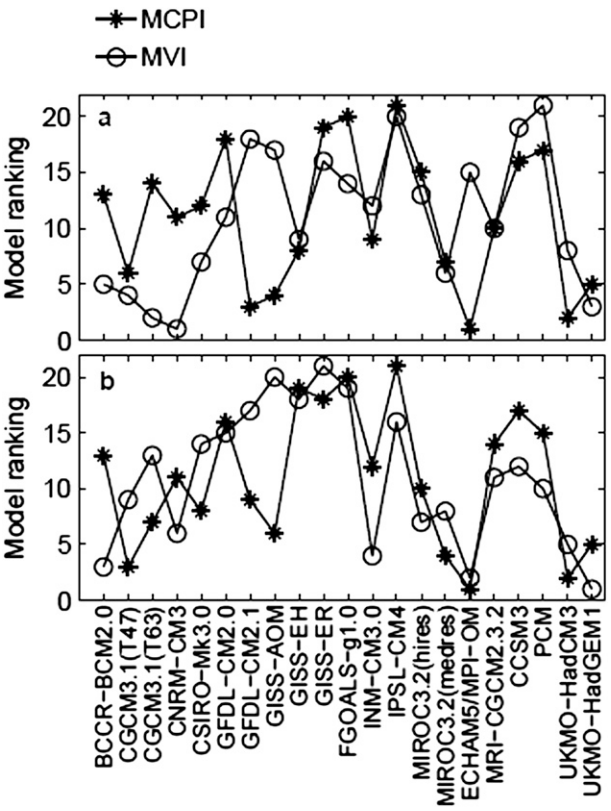

FIG. 7. Model ranking according to MCPI and MVI: (a) small domain and (b) large domain.

Is skill in simulating frequencies of daily SLP patterns related to skill in simulating the mean annual cycle and the interannual variability of SLP? We address this question by comparing the model ranking according to three measures: RMSE [Eq. (1)], variance ratio [$\beta^2 - 1/\beta^2$ in Eq. (2)], and M_R measure calculated for temporal and for spatial SLP anomaly patterns over all the seasons and the three SOM sizes (Fig. 9). Correlating the model rankings over the small domain (Fig. 9a), we find that the only significantly positive correlation ($r = 0.68$) is for the skill in simulating the mean annual cycle of SLP and the frequency of the spatial SLP anomaly patterns. Over the large domain, the only significantly positive correlation ($r = 0.56$) is between the ranking for the mean annual cycle of SLP and the ranking for the skill in simulating the frequencies of the temporal SLP anomaly patterns.

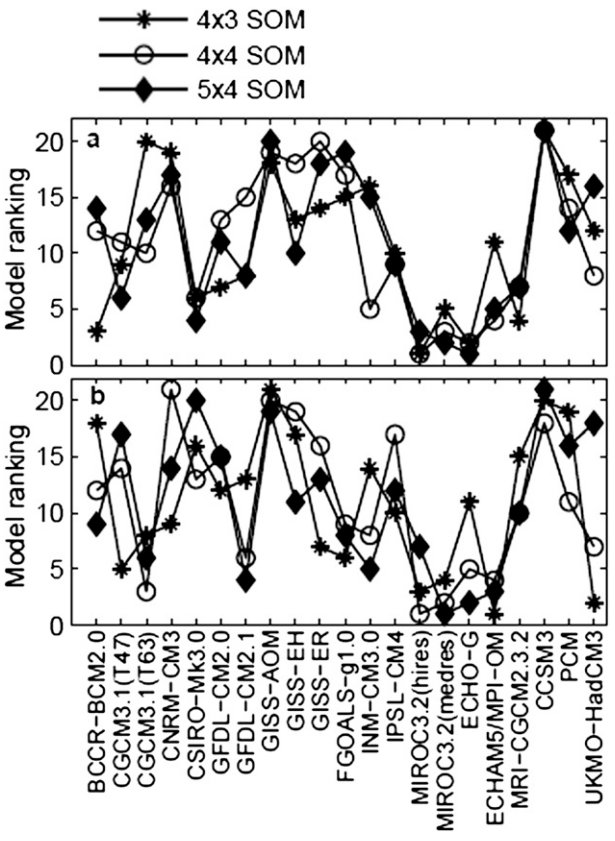

FIG. 8. Model ranking according to the sum of measures M_C and M_S assessed for three different SOM sizes: 4×3, 4×4, and 5×4: (a) small domain and (b) large domain.

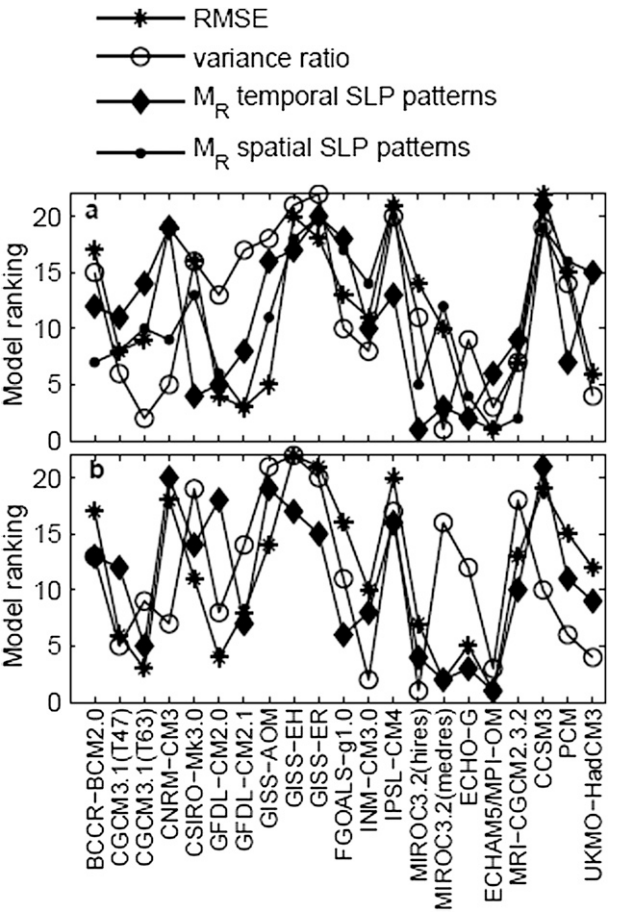

FIG. 9. Model ranking according to the RMSE for SLP, variance ratio for SLP, and the rank measure calculated for temporal and spatial SLP anomaly patterns over all the seasons and the three SOM sizes: (a) small domain and (b) large domain.

Figure 10 summarizes the model ranking from the entire set of evaluation measures we used in this study, all assessed over the two baseline periods 1970–89 and 1980–99. Relating each rank to a color (Fig. 10), the scatter of colors illustrates that the model ranks vary considerably across the measures. However, some models consistently rank high {e.g., ECHAM5–MPI-OM and the Model for Interdisciplinary Research on Climate 3.2, medium-resolution version [MIROC3.2(medres)]} and some consistently rank low (e.g., CCSM3 and GISS models). Summing up all the ranks for each model, our top-five models are ECHAM5–MPI-OM, MIROC3.2(medres), MIROC3.2(hires), CGCM3.1(T47), and CGCM3.1(T63). Although ECHO-G scores high, its evaluation has been performed over only two available climate variables (PR and SLP). Similarly, HadGEM scores high according to the statistical measures, but unavailability of daily data prevented the evaluation using SOMs.

Figure 10 also reveals that the GCMs developed by the same modeling center (three GISS models, and a pair of models in GFDL, MIROC, CGCM, and UKMO) tend to rank near each other. We further test this observation by assessing the differences in model rank for each model pair from the sample of 22 GCMs across all evaluation metrics. A median rank difference in this sample is 7, and we refer to it as a typical model difference. Taking only the rank differences of model pairs from the same modeling center, we find that 68% of differences are smaller than the typical model difference. Some models from the same center differ only in their resolution (e.g., CGCM and MIROC) and therefore their difference in performance might be expected to be smaller than across models from different centers.

In sections 4a–c we showed the sensitivity of evaluation results to the choice of the two baseline periods. Here we reanalyze our findings by calculating the correlation between the model ranking for the period 1970–89 and 1980–99 for each evaluation measure in Fig. 10. All r values are found to be significantly positive at 95% confidence level, with the highest value of $r = 0.98$ when the ranking according to RMSE for SLE over the large

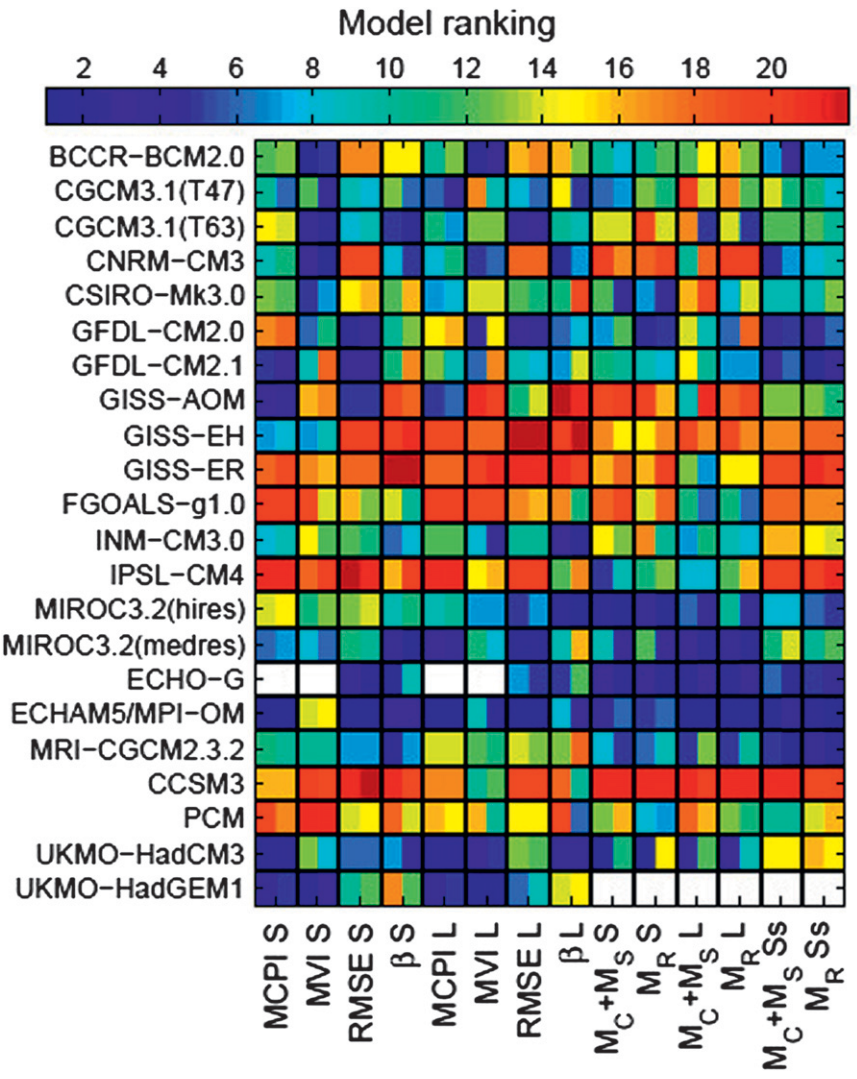

FIG. 10. Model ranking (from best performing with rank = 1 to worst performing model with rank = 22) according to the set of metrics over the small (S) and large (L) domain. The metrics are MCPI, MVI, RMSE for SLP, variance ratio β^2 for SLP, sum of metrics M_C and M_S, and metric M_R, all calculated over all the seasons and all three SOM sizes for temporal (t) and spatial (s) SLP anomaly patterns. White squares indicate the unavailability of model data. Two colors per square correspond to the ranking performed over the two baseline periods (left) 1970–89 and (right) 1980–99.

domain is used, and the lowest value of $r = 0.60$ when the ranking according to variance ratio for SLE over the small domain is used. We conclude that the sensitivity of the model ranking in Fig. 10 to the choice of the two baseline periods is small.

In Table 3 we compare our list of the top five models with the lists of top-performing models from some recent GCM evaluations over different regions and climate variables of interest. Some of these studies used statistical measures of evaluation (RMSE, variance ratios) while some used SOMs and the correlations between node frequencies. As illustrated in the table, some top-performing models appear on different lists, but only ECHAM5–MPI-OM appears in every list. Additionally, the worst-performing models in our study do not appear in any list with the five top-performing models. The only exception is CCSM3, which showed good performance in simulating the frequency of synoptic patterns over Greenland (Schuenemann and Cassano 2009) while failing to perform well in our study. One of the possible reasons for different model rankings between our studies and others is the choice of the reference data. Our study is the only one that used the NARR for the reference climate whereas for most of those listed in Table 3 used the 40-yr European

TABLE 3. The top-ranking models from different studies for different domains and based on different evaluation metrics. Ranking in our study is based on model performance across all the metrics in the text.

	Rank				
	1	2	3	4	5
This study	ECHAM5–MPI-OM	MIROC3.2(medres)	MIROC3.2(hires)	CGCM3.1(T47)	CGCM3.1(T63)
Gleckler et al. (2008)					
20°–90°N	HadGEM1	ECHAM5/MPI-OM	HadCM3	GFDL CM2.1	MIROC3.2(hires)
Walsh et al. (2008)					
Alaska	GFDL CM2.0	GFDL CM2.1	HadCM3		
Greenland	GFDL CM2.1	ECHAM5/MPI-OM	MIROC3.2(medres)		
60°–90°N	ECHAM5–MPI-OM	GFDL CM2.1	MIROC3.2(medres)		
20°–90°N	ECHAM5–MPI-OM	GFDL CM2.1	MIROC3.2(medres)	and CGCM3.1(T63)	
Finnis et al. (2008)					
Mackenzie River basin	ECHAM5–MPI-OM	GFDL CM2.1	CGCM3.1(T63)	MIROC3.2(hires)	
Schuenemann and Cassano (2009)					
Greenland	MIROC3.2(hires)	CGCM3.1(T63)	ECHAM5–MPI-OM	GFDL CM2.1	CCSM3

Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40).

Why do some models consistently score high [e.g., ECHAM5–MPI-OM, GFDL CM2.1, MIROC3.2(medres) in Table 3] independent of the evaluation measures, region, and variable of interest? One reason might be the model resolution, yet we find no systematic relationship between our model ranking and the resolution of the GCMs (see also Walsh et al. 2008). Nevertheless, according to most of the evaluation measures in our analysis, the models with the coarsest resolution score low [all GISS models and Institute of Numerical Mathematics Coupled Model, version 3.0 (INM-CM3.0)]. There is also no systematic relationship between model performance and whether or not flux adjustment has been applied in the GCM [i.e., adjustment of the surface heat, water, and momentum fluxes in order to maintain a stable control climate; Table 8.1 in Randall et al. (2007) provides details as to which GCMs apply the flux adjustments]. Other possible reasons for the differences in model performance include the cloud and radiative formulations, the planetary boundary layer parameterizations, and the land surface schemes. Specifically, different levels of model performance over our region of interest (small and large domains) might depend on the biases in the large-scale atmospheric circulation driven by processes outside the region.

5. Summary and conclusions

Using a set of evaluation measures we have analyzed the performance of 22 GCMs over all North America and its western subregion. Emphasis has been given to the evaluation of the model skill in simulating climatologies of several climate variables and the characteristic synoptic patterns of daily sea level pressure. The reference data to which the modeled climate fields have been compared was the North American Regional Reanalysis over the period 1980–99. The better the agreement with the NARR, the higher is the score of the model.

Statistical measures were used to compare modeled versus observed mean annual cycle and interannual variability of 6 selected variables: precipitation, sea level pressure, geopotential height at 850 hPa and 500 hPa, specific humidity at 850 hPa, and air temperature at 850 hPa. According to these measures some models score better than others, although no model scores above or below average for all six climate variables. Models that perform well over all North America (the large domain) also perform well over its western subregion (the small domain) for all climate variables except precipitation. Model skill is interrelated among some variables. For example, skill in simulating the mean annual cycle and interannual variability of sea level pressure is correlated with skill in simulating the same features for geopotential height at 850 hPa. Some correlations between skill across the climate variables are shown to be sensitive to the size of the domain. We find weak or insignificant relationship between the ability to simulate the mean climate and the ability to simulate the interannual variability, supporting the results from Gleckler et al. (2008). Our results also support the previous findings (e.g., Gleckler et al. 2008; Pierce et al. 2009) that the multimodel ensemble average is superior to any individual model in simulating the mean annual cycle. Based on statistical metrics the five top performing models are ECHAM5–MPI-OM, HadCM3, CGCM3.1(T47), MIROC3.2(medres), and CGCM3.1(T63).

To provide another set of evaluation measures, we applied the method of self-organizing maps (SOMs) to identify and classify the characteristic daily patterns of sea level pressure (SLP) over the region. The SOM

nodes represent archetypal patterns derived for temporal and spatial SLP anomalies over the small and large domain and for each individual season. The evaluation consisted of correlating the frequencies of these patterns, as generated in GCMs, to the frequencies in the NARR. For the SOM nodes of temporal SLP anomalies, which show the 2D distribution of biases from the 20-yr average of daily SLP pattern, most of the models have significantly positive correlations over all 4 seasons. This is because most of the models are able to simulate well the 20-yr average of daily SLP pattern over the region. Despite this success in model performance, very few GCMs are able to reproduce the frequencies of the spatial SLP anomaly patterns in the NARR on seasonal basis over the baseline period. Many models with under-average performance in winter overestimate the occurrence of patterns with a strong Aleutian low. In summer, poor-performing models underestimate the occurrence of a low pressure center over Arctic North America in conjunction with the northeastern Pacific subtropical anticyclone. The number of significantly positive correlations over all four seasons between the modeled and the NARR frequencies differs for the small and large domains. If one ignores correlation significance and ranks the models from best to worst based on their correlation with the NARR frequencies, the overall model ranking over the small domain is very similar to the ranking over the large domain. Likewise, the ranking for temporal SLP patterns is similar to the ranking for spatial SLP patterns. Thus, the choice of evaluation measures depends on whether one is interested only in the relative model performance or in the ability to simulate the climate features of interest. Considering both absolute and relative model performance in this SOM analysis over the small and large domain, the five top-performing models are MRI-CGCM2.3.2, ECHAM5–MPI-OM, MIROC3.2(hires), ECHO-G, and BCCR.

Our results demonstrate that even with a targeted set of climate variables and domain of interest, the skill of a model remains sensitive to the choice of the climate variable and to the size of spatial domain. Additionally, the skill of the model in simulating a climate feature that displays interdecadal oscillation is sensitive to the 10-yr shift in the 20-yr baseline period over which the skill is assessed. In our analysis, this sensitivity is particularly well illustrated for the simulations of temporal SLP anomaly patterns in winter, when the predominant Aleutian low pattern experiences a regime shift as part of Pacific interdecadal variability.

Inconsistency in the model skill to simulate a broad spectrum of climate features shows the need to identify the features that are the most important for the intended application. For modeling the deglaciation of North America, the application we have in mind, the choice of the GCMs will depend on the selected downscaling method. For example, if we used a method of "weather typing" for statistical downscaling of GCMs (e.g., Zorita and von Storch 1999; Enke et al. 2005) the logical candidates would be the top-performing models according to the evaluation using SOMs. If we chose to combine different statistical downscaling methods (e.g., linear methods and neural networks) we would select the models that score high over the whole set of measures and climate variables.

Despite the large scatter of model performance, our results reveal that some models consistently rank high across all the evaluation metrics in our region [e.g., ECHAM5–MPI-OM and MIROC3.2(medres)]. Some models consistently rank low [e.g., CCSM3 and GISS Model E-H (GISS-EH)]. Our selection of GCMs relies on the assumption that models that best simulate present-day climate also yield the best projections of future climate. It remains uncertain how model bias in the present transfers into different projections of the future.

Acknowledgments. We thank C. Reuten, A. Jarosch, F. Anslow, B. Ainslie, D. Moore, F. Weber, and S. Fleming for their constructive discussions and valuable criticism. G. Flato generously provided the NARR data. Furthermore, we thank A. Rasmussen, A. Werner, and the three anonymous reviewers for their valuable comments. Financial support for this project was provided through the Polar Climate Stability Network and the Western Canadian Cryospheric Network, both funded by the Canadian Foundation for Climate and Atmospheric Sciences, and by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Ambroise, C., G. Seze, F. Badran, and S. Thiria, 2000: Hierarchical clustering of self-organizing maps for cloud classification. *Neurocomputing,* **30,** 47–52.

Arendt, A., J. Walsh, and W. Harrison, 2009: Changes of glaciers and climate in northwestern North America during the late twentieth century. *J. Climate,* **22,** 4117–4134.

Barnett, T. P., and Coauthors, 2008: Human-induced changes in the hydrology of the western United States. *Science,* **319,** 1080–1083.

Bitz, C. C., and D. S. Battisti, 1999: Interannual to decadal variability in climate and the glacier mass balance in Washington, western Canada, and Alaska. *J. Climate,* **12,** 3181–3196.

Brekke, L. D., M. D. Dettinger, E. P. Maurer, and M. Anderson, 2008: Significance of model credibility in estimating climate projection distributions for regional hydroclimatological risk assessments. *Climatic Change,* **89,** 371–394.

Cavazos, T., 2000: Using self-organizing maps to investigate extreme climate events: An application to wintertime precipitation in the Balkans. *J. Climate,* **13,** 1718–1732.

Coquard, J., P. B. Duffy, and K. E. Taylor, 2004: Present and future surface climate in the western U.S. as simulated by 15 global climate models. *Climate Dyn.,* **23,** 455–472.

Dettinger, M. D., 2005: From climate change spaghetti to climate-change distributions for 21st century California. *San Francisco Estuary Watershed Sci.,* **3,** 1–14.

Enke, W., F. Schneider, and T. Deuschländer, 2005: A novel scheme to derive optimized circulation pattern classifications for downscaling and forecast purposes. *Theor. Appl. Climatol.,* **82,** 51–63.

Finnis, J., J. Cassano, M. Holland, M. Serreze, and P. Uotilla, 2008: Synoptically forced hydroclimatology of major Arctic watersheds in general circulation models. Part 1: The Mackenzie River Basin. *Int. J. Climatol.,* **29** (9), 1226–1243.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.,* **113,** L06711, doi:10.1029/2007JD008972.

Hewitson, B. C., and R. G. Crane, 1994: *Neural Nets: Applications in Geography*. Springer, 208 pp.

——, and ——, 2002: Self-organizing maps: Applications to synoptic climatology. *Climate Res.,* **22,** 13–26.

Hodge, S. M., D. C. Trabant, R. M. Krimmel, T. A. Heinrichs, R. M. March, and E. G. Josberger, 1998: Climate variations and changes in mass of three glaciers in western North America. *J. Climate,* **11,** 2161–2179.

Hong, Y., K. Hsu, S. Sorooshian, and X. Gao, 2004: Precipitation estimation from remotely sensed imagery using an artificial neural network cloud classification system. *J. Appl. Meteor.,* **43,** 1834–1853.

——, ——, ——, and ——, 2005: Self-organizing non-linear output (SONO): A neural network suitable for cloud patch-based rainfall estimation at small scales. *Water Resour. Res.,* **41,** W03008, doi:10.1029/2004WR003142.

Hsu, K., H. V. Gupta, X. Gao, S. Sorooshian, and B. Imam, 2002: Self-organizing linear output map (SOLO): An artificial neural network suitable for hydrologic modeling and analysis. *Water Resour. Res.,* **38,** 1302, doi:10.1029/2001WR000795.

Jarosch, A. H., F. S. Anslow, and G. K. C. Clarke, 2010: High resolution precipitation and temperature down-scaling for glacier models. *Climate Dyn.,* doi:10.1007/s00382-010-0949-1.

Kaski, S., J. Kangas, and T. Kohonen, 1998: Bibliography of self-organizing map (SOM) papers: 1981–1997. *Neural Comput. Surv.,* **1,** 102–350.

Kohonen, T., 2000: *Self-Organizing Maps*. 3rd ed. Springer, 528 pp.

Liu, Y., R. H. Weisberg, and C. N. K. Mooers, 2006: Performance evaluation of the Self-Organizing Map for feature extraction. *J. Geophys. Res.,* **111,** C05018, doi:10.1029/2005jc003117.

Malmgren, B. A., and A. Winter, 1999: Climate zonation in Puerto Rico based on principal components analysis and an artificial neural network. *J. Climate,* **12,** 977–985.

Matulla, C., E. Watson, S. Wagner, and W. Schöner, 2008: Downscaled GCM projections of winter and summer mass balance for Peyto Glacier, Alberta, Canada (2000-2100) from ensemble simulations with ECHAM5-MPIOM. *Int. J. Climatol.,* **29,** 1550–1559.

McCabe, G. J., and A. G. Fountain, 1995: Relations between atmospheric circulation and mass balance of South Cascade Glacier, Washington, U.S.A. *Arct. Alp. Res.,* **27,** 226–233.

Mesinger, F., and Coauthors, 2006: North American Regional Reanalysis. *Bull. Amer. Meteor. Soc.,* **87,** 343–360.

Milly, P. C. D., K. A. Dunne, and A. V. Vecchia, 2005: Global pattern of trends in streamflow and water availability in a changing climate. *Nature,* **438,** 347–350.

Oja, M., S. Kaski, and T. Kohonen, 2002: Bibliography of Self-Organizing Map (SOM) papers: 1998–2001 addendum. *Neural Comput. Surv.,* **3,** 1–156.

Pierce, D. W., T. P. Barnett, B. D. Santer, and P. J. Gleckler, 2009: Selecting global climate models for regional climate change studies. *Proc. Natl. Acad. Sci. USA,* **106,** 8441–8446, doi:10.1073/pnas.0900094106.

Pincus, R., C. P. Batstone, R. J. P. Hofmann, K. E. Taylor, and P. J. Gleckler, 2008: Evaluating the present-day simulation of clouds, precipitation, and radiation in climate model. *J. Geophys. Res.,* **113,** D14209, doi:10.1029/2007JD009334.

Randall, D. A., and Coauthors, 2007: Climate models and their evaluation. *Climate Change 2007: The Physical Science Basis,* S. Solomon et al, Eds., Cambridge University Press, 589–662.

Rasmussen, L. A., and H. Conway, 2001: Estimating South Cascade Glacier mass balance from a distant radiosonde and comparison with Blue Glacier. *J. Glaciol.,* **47,** 579–588.

——, and ——, 2003: Using upper-air conditions to estimate South Cascade Glacier (Washington, U.S.A.) summer balance. *J. Glaciol.,* **49,** 456–462.

Reichler, T., and J. Kim, 2008: How well do coupled models simulate today's climate? *Bull. Amer. Meteor. Soc.,* **89,** 303–311.

Reusch, D. B., R. B. Alley, and B. C. Hewitson, 2005: Relative performance of self-organizing maps and principal component analysis in pattern extraction from synthetic climatological data. *Polar Geogr.,* **29,** 227–251.

Schuenemann, K. C., and J. J. Cassano, 2009: Changes in synoptic weather patterns and Greenland precipitation in the 20th and 21st centuries: 1. Evaluation of late 20th century simulations from IPCC models. *J. Geophys. Res.,* **114,** D20113, doi:10.1029/2009JD011705.

Shea, J. M., and S. J. Marshall, 2007: Atmospheric flow indices, regional climate, and glacier mass balance in the Canadian Rocky Mountains. *Int. J. Climatol.,* **27,** 233–247, doi:10.1002/joc.1398.

Tangborn, W. V., 1980: Two models for estimating climate glacier relationships in the North-Cascades, Washington, U.S.A. *J. Glaciol.,* **25,** 3–21.

Taylor, K. E., and P. J. Gleckler, 2002: The Second Phase of the Atmospheric Model Intecomparison Project. Lawrence Livermore National Laboratory Rep. Series, Rep. UCRL-PROC-209115, 253 pp.

Tebaldi, C., and R. Knutti, 2007: The use of the multimodel ensemble in probabilistic climate projections. *Philos. Trans. Roy. Soc. London,* **365A,** 2053–2075.

——, R. L. Smith, D. Nychka, and L. O. Mearns, 2005: Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. *J. Climate,* **18,** 1524–1540.

Trenberth, K. E., and J. W. Hurrell, 1994: Decadal atmosphere-ocean variations in the Pacific. *Climate Dyn.,* **9,** 303–319.

Vesanto, J., J. Himberg, E. Alhoniemi, and J. Parhankangas, 2000: SOM toolbox for Matlab 5. Helsinki University of Technology, Rep. A57, 59 pp.

Walsh, J. E., W. L. Chapman, V. E. Romanovsky, J. H. Christensen, and M. Stendel, 2008: Global climate model performance over Alaska and Greenland. *J. Climate,* **21,** 6156–6174.

Wang, M., and J. E. Overland, 2009: A sea ice free summer Arctic within 30 years? *Geophys. Res. Lett.,* **36,** L07502, doi:10.1029/2009GL037820.

Yarnal, B., 1984: Relationships between synoptic-scale atmospheric circulation and glacier mass balance in southwestern Canada during the International Hydrological Decade, 1963–74. *J. Glaciol.,* **30,** 188–198.

Zorita, E., and H. von Storch, 1999: The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *J. Climate,* **12,** 2474–2489.