Storm Track Predictability on Seasonal and Decadal Scales

GILBERT P. COMPO AND PRASHANT D. SARDESHMUKH

NOAA–CIRES Climate Diagnostics Center, University of Colorado, Boulder, Colorado

(Manuscript received 17 September 2003, in final form 9 April 2004)

ABSTRACT

This paper is concerned with estimating the predictable variation of extratropical daily weather statistics (''storm tracks'') associated with global sea surface temperature (SST) changes on interannual to interdecadal scales, and its magnitude relative to the unpredictable noise. The SST-forced storm track signal in each northern winter in 1950–99 is estimated as the mean storm track anomaly in an ensemble of atmospheric general circulation model (AGCM) integrations for that winter with prescribed observed SSTs. Two sets of ensembles available from two modeling centers, with anomalous SSTs prescribed either globally or only in the Tropics, are used. Since the storm track signals cannot be derived directly from the archived monthly AGCM output, they are diagnosed from the SST-forced winter-mean 200-mb height signals using an empirical linear storm track model (STM). For two particular winters, the El Niño of January–February–March (JFM) 1987 and the La Niña of JFM 1989, the storm track signals and noise are estimated directly, and more accurately, from additional large ensembles of AGCM integrations. The linear STM is remarkably successful at capturing the AGCM's storm track signal in these two winters, and is thus also suitable for estimating the signal in other winters.

The principal conclusions from this analysis are as follows. A predictable SST-forced storm track signal exists in many winters, but its strength and pattern can change substantially from winter to winter. The correlation of the SST-forced and observed storm track anomalies is high enough in the Pacific–North America (PNA) sector to be of practical use. Most of the SST-forced signal is associated with tropical Pacific SST forcing; the central Pacific (Niño-4) is somewhat more important than the eastern Pacific (Niño-3) in this regard. Variations of the pattern correlation of the SST-forced and observed storm track anomaly fields from winter to winter, and among five-winter averages, are generally consistent with variations of the signal strength, and to that extent are identifiable a priori. Larger pattern correlations for the five-winter averages found in the second half of the 50-yr record are consistent with the stronger El Niño SST forcing in the second half. *None of these conclusions, however, apply in the Euro-Atlantic sector,* where the correlations of the SST-forced and observed storm track anomalies are found to be much smaller. Given also that they are inconsistent with the estimated signal-to-noise ratios in this region, substantial AGCM error in representing the regional response to tropical SST forcing, rather than intrinsically lower Euro-Atlantic storm track predictability, is argued to be behind these lower correlations.

1. Introduction

It is well known that the statistics of extratropical daily weather (''storm tracks'') averaged over individual winter seasons, decades, or even longer intervals are not constant but vary substantially from one interval to the next. These variations have a random part associated with sampling fluctuations, and a potentially predictable part associated with slow changes of atmospheric boundary conditions and atmospheric composition. This paper addresses the problem of estimating the predictable signal associated specifically with sea surface temperature (SST) changes, and its magnitude relative to the random noise.

The ratio S of the predictable signal of any quantity (such as a storm track anomaly) to its random noise has a simple relationship to the expected correlation of the predicted and observed values of that quantity (e.g., Sardeshmukh et al. 2000), and is thus a useful measure of potential predictability. As discussed in the appendix, the expected correlation skill ρ_n of an n-member ensemble-mean forecast made by a ''perfect'' model is

$$\rho_n = S^2/[(S^2 + 1)(S^2 + n^{-1})]^{1/2}. \quad (1)$$

The thin curves in Fig. 1 illustrate this relationship for a few values of n. The outermost ρ_∞ curve shows how predictability is limited if S is small; this limitation cannot be overcome even using infinite-member ensembles of a perfect model. The expected skill ρ_n using n-member ensembles is lower than ρ_∞, and model error (see appendix) leads to even lower actual skill ρ. In this framework, the problem of estimating predictability becomes essentially one of estimating S. As illustrated by the thickened portions of the curves, using n-member ensembles introduces errors in estimating S, and hence in estimating ρ_n. For low n this uncertainty in ρ_n is much

Corresponding author address: Dr. Gilbert P. Compo, NOAA–CIRES Climate Diagnostics Center R/CDC1, 325 Broadway, Boulder, CO 80305-3328.
E-mail: compo@colorado.edu

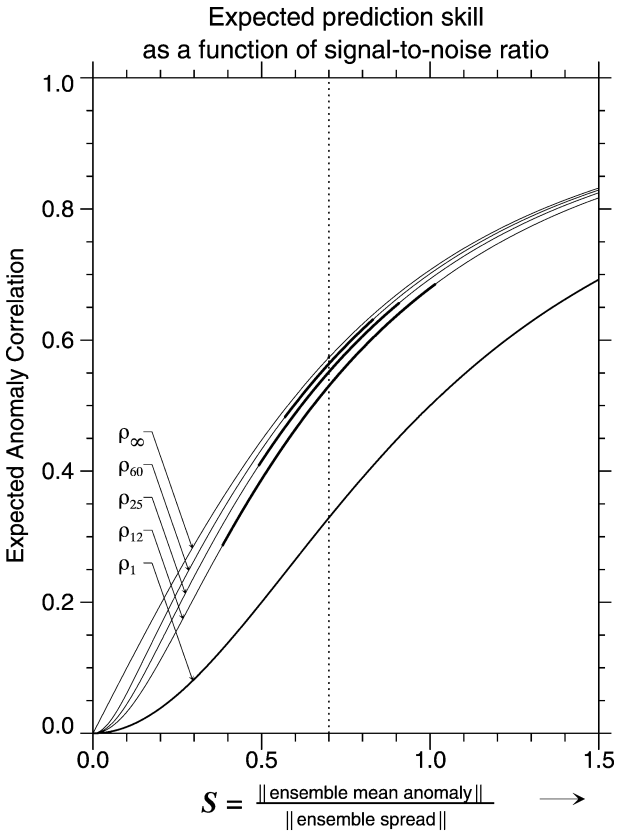

FIG. 1. Expected anomaly correlation skill ρ_n of ensemble-mean forecasts for $n = 1$, 12, 25, 60, and infinite member ensembles as a function of the signal-to-noise ratio S [Eq. (1)]. Thickened portions of curves illustrate uncertainty in expected skill ρ_n for $S = 0.7$ due to uncertainty from estimating S using an n-member ensemble, assuming that S is distributed as a Student's t statistic. (Adapted from Sardeshmukh et al. 2000.)

larger than the difference between ρ_∞ and ρ_n, and can lead to spuriously low (or high) predictability estimates.

With these considerations in mind, we will examine the evidence for potentially predictable storm track signals. Observational studies suggest the existence of an interannual storm track signal arising from SST changes associated with the El Niño–Southern Oscillation (ENSO) phenomenon. Previous work has shown that this signal extends eastward from the central North Pacific across North America and the Atlantic to Europe (e.g., Fraedrich and Muller 1992; Fraedrich 1994; Hoerling and Ting 1994; Straus and Shukla 1997; May and Bengtsson 1998; Matthews and Kiladis 1999; Smith and Sardeshmukh 2000; Sardeshmukh et al. 2000; Carillo et al. 2000; Compo et al. 2001). The observations also suggest somewhat different effects for El Niño and La Niña forcing, thus hinting that the signal may vary from ENSO case to case, with obvious implications for storm track predictability. The limited observational record, however, compromises estimating such case-dependent and/or nonlinear signals with statistical significance, and also compromises estimating the storm track noise essential for assessing predictability. A similar remark applies to most previous assessments of these effects made using small atmospheric general circulation model (AGCM) ensembles. To remedy this situation, Compo et al. (2001) examined much larger 60-member ensembles of seasonal integrations of the National Centers for Environmental Prediction (NCEP) AGCM with prescribed observed global SSTs for one El Niño [January–February–March (JFM) 1987] and one La Niña (JFM 1989) case, and were able to conclude with much greater confidence that the SST-forced storm track signal may indeed vary substantially from case to case, especially over the North Atlantic and Europe. Interestingly, they could also demonstrate a statistically significant storm track signal in many regions not usually associated with an ENSO effect.

Demonstrating the existence of a signal is, of course, not the same as demonstrating its predictability, or usefulness. As illustrated in Fig. 1, what matters for predictability and usefulness is the size of the signal relative to the noise. Compo et al. did not consider this question explicitly; it is our principal concern here.

The signal-to-noise ratio S for any quantity may be estimated from ensemble integrations as the ratio of the ensemble-mean anomaly to the ensemble spread. Figure 2b shows S for the SST-forced 500-mb vertical velocity (omega) storm track anomalies in JFM 1987 estimated from the 60-member NCEP AGCM ensembles considered by Compo et al. To generate this plot, an ensemble member's winter storm track value was defined at each grid point, as in Compo et al., as the 2–7-day bandpass-filtered variance of 500-mb omega. Storm track anomalies for each member were computed with respect to the ensemble-mean storm track in a separate 90-member ensemble with climatological SST boundary forcing. The storm track signal associated with anomalous JFM 1987 SSTs was then computed as the ensemble mean of the 60 storm track anomaly values, and the noise as the rms deviation of the 60 values from this mean. Figure 2 also shows the S values for the SST-forced winter-mean 500-mb omega and precipitation anomalies calculated in a similar manner. As suspected, the storm track S values are modest, but comparable in magnitude to the S values for 500-mb omega and precipitation. The figure nevertheless strongly suggests that the predictability of winter-mean precipitation is as much tied to the predictability of the winter-mean 500-mb omega storm track as it is to that of the winter-mean 500-mb omega. Mainly for this reason, we will restrict ourselves to the predictability of the ''500-mb omega storm track'' in this paper, as opposed to many other interesting measures of synoptic variability.

The modest values of S for the omega storm track in Fig. 2 apparently imply only modest storm track predictability associated with SST changes. It is important to keep in mind, however, that these estimates are based on one specific winter case and one particular AGCM. It is unclear to what extent they are affected by the

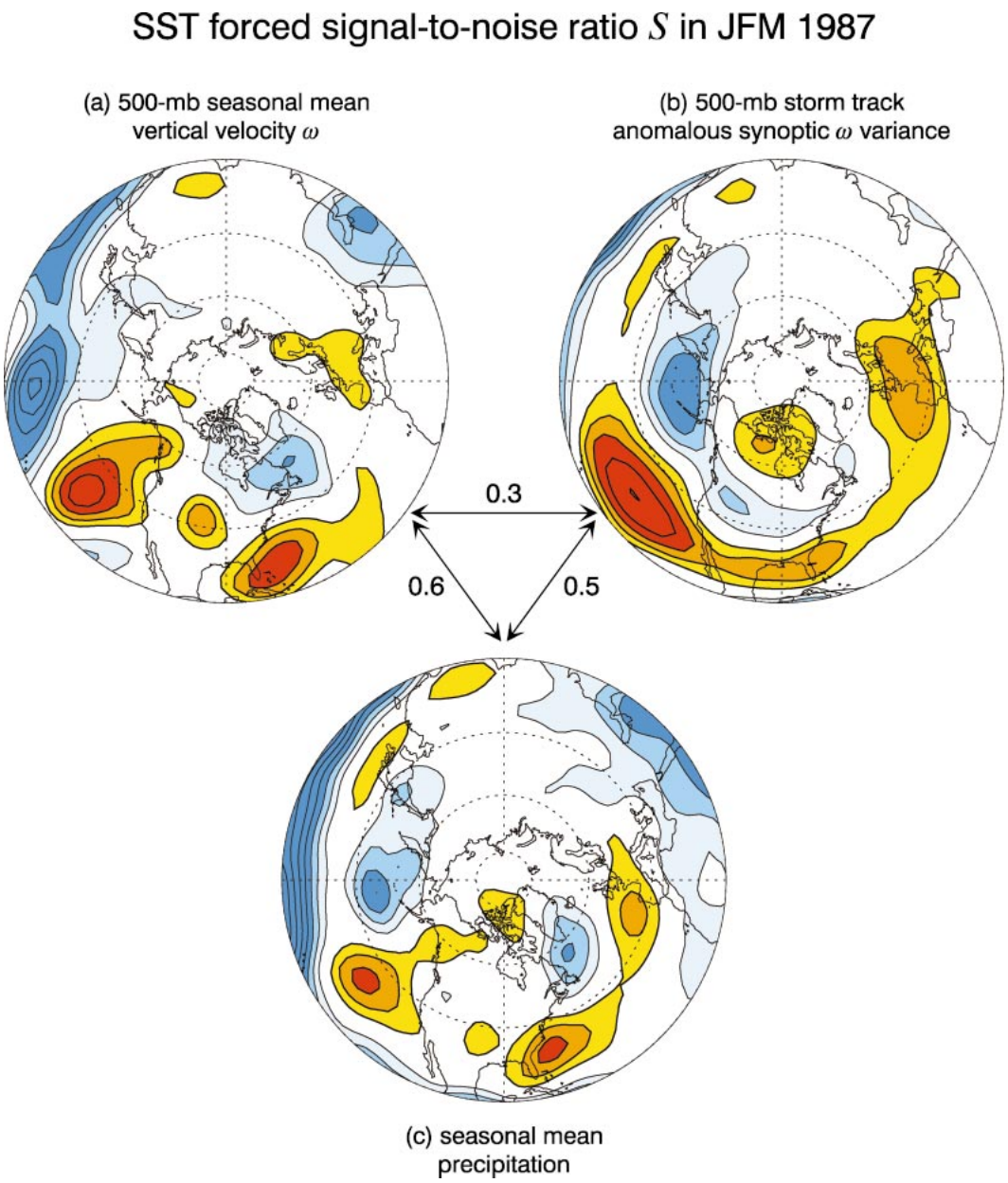

FIG. 2. Signal-to-noise ratio *S* from the JFM 1987 El Niño for (a) seasonal-mean 500-mb vertical velocity, (b) seasonal 2–7-day bandpass variance of vertical velocity (ω), and (c) seasonal-mean precipitation. The contour interval is 0.2. The zero contour has been suppressed. Thick (thin) contours indicate positive (negative) values. The 10% significance level is 0.22 using a two-sided *t* test. All plots are field significant at the 5% level assuming at least 15 esdof. Pattern correlations between the respective fields are indicated next to the arrows.

specifics of that case and/or model error. Comparing the AGCM's ensemble-mean predicted storm track anomaly with the *observed* anomaly in JFM 1987 (not shown) does not settle the issue, because the AGCM's prediction is only the *expected* anomaly. The prediction problem is inherently probabilistic, so the reliability of model-generated predictability estimates can only be assessed by examining prediction skill over a large number of cases.

To make a more general assessment of wintertime storm track predictability, we will therefore compare predictability estimates obtained using Eq. (1) with actual correlations of the SST-forced and observed storm track variations over the last half century. To this end, we will use two sets of relatively small (9 to 13 members) ensemble runs for the last half century available from two modeling centers [NCEP and National Center for Atmospheric Research (NCAR)], with anomalous observed SSTs prescribed either globally or only in the Tropics, to estimate the SST-forced signal in each winter.

TABLE 1. Details and time periods of the integrations of atmospheric general circulation models available at twice-daily and monthly resolution used in the present study. Here, T refers to the spectral truncation of the model, and L refers to the number of model levels. T40 (T42) corresponds to a spatial resolution of approximately 3° lat × 3° lon (2.8° × 2.8°).

Model	Boundary condition	Time period	No. of members	Reference
NCEP MRF9 T40L18 (570 JFM, twice daily)	Global SSTs	Climatological	90	Sardeshmukh et al. (2000)
		Jan–Mar 1987	60	
		Jan–Mar 1989	60	Compo et al. (2001)
	Niño-4 anomaly	±1 Jan–Mar	90 each	This study
		±3, ±5 Jan–Mar	45 each	
NCEP MRF9 T40L18 (AMIP style, monthly)	Global SSTs (GOGA)	1950–95	13	Livezey et al. (1997)
	Pacific SST anomaly (POGA)	1950–95	9	Kumar and Hoerling (1998)
NCAR CCM3 T42L18 (AMIP style, monthly)	Global SSTs (GOGA)	1950–99	12	Kiehl et al. (1998)
	Tropical SST anomaly (TOGA)	1950–99	11	

Since the storm track signals cannot be obtained directly from the archived monthly AGCM output, we will diagnose them from the winter-mean 200-mb-height signals using an empirical linear storm track model (STM) developed specifically for this purpose. We will present the correlations ρ of these STM-diagnosed storm track signals with the observed storm track anomalies, both as maps of the temporal correlation of the SST-forced and observed anomaly values over 50 winters at each grid point, and as 50-winter time series of the pattern correlation of the SST-forced and observed anomaly fields in each winter over the Pacific–North American (PNA) and North Atlantic–Europe (NATL-EUR) sectors. The latter, in comparison with the time series of tropical SST indices, will help us assess the case dependence of storm track predictability.

Our main interest in this paper is in extracting the predictable component of storm track variations. The words *predictable* and *SST forced* may be used almost interchangeably for interannual storm track variations, given the importance of the potentially predictable interannual tropical SST variations in forcing them. Several studies have also found substantial decadal storm track variations and trends over the last 50–100 yr (e.g., Hurrell and van Loon 1997; WASA Group 1998; Graham and Diaz 2001; Chang and Fu 2002, 2003; Gulev et al. 2002; Harnik and Chang 2003). The decadal variations of the omega storm track, with its more direct link to precipitation variations noted earlier, have not been previously studied, and the degree to which they are SST-forced has also not been addressed. Acknowledging that the existence of an SST-forced component does not by itself imply storm track predictability on decadal scales, we will nevertheless also present correlations of five-winter averages of the SST-forced and observed storm track anomalies, and explore to what extent they are associated with anomalous five-winter-average tropical SSTs.

The paper is organized as follows. The data and model integrations are discussed in section 2. In section 3, the linear STM is developed and tested for its ability to reproduce the NCEP AGCM's (60 member) SST-forced storm track signals in JFM 1987 and 1989, given only the AGCM's ensemble-mean 200-mb-height signals in those winters. In section 4, the STM is used to diagnose the SST-forced storm track signals in 1950–99, given the NCEP and NCAR AGCMs' ensemble-mean 200-mb-height responses to observed global and tropical SST forcing. The AGCMs' skill in simulating the observed storm track anomalies is then evaluated through the correlation measures discussed earlier. This average skill is compared with that expected from Eq. (1) for the 1987 and 1989 events to illustrate the case dependence of expected forecast skill. In section 5, the STM is used to address the important issue of decadal storm track variations. In section 6, we show that the actual AGCM skill in predicting storm tracks is close to the expected skill over the Pacific–North American sector, but that a substantial systematic error may be present over the North Atlantic–European sector. A discussion and concluding remarks follow in section 7.

2. Data

This section provides details of the data sources and data analysis procedures to facilitate the reproducibility of our results by other investigators. Those not interested in such details may proceed to section 3 without loss.

Both observational and AGCM datasets were used in our analysis. Observed geopotential height and vertical velocity fields were obtained from 50 yr (1950–99) of NCEP–NCAR reanalyses (Kistler et al. 2001) at twice-daily resolution. SST indices for the Niño-3 (5°N–5°S, 150°–90°W) and Niño-4 (5°N–5°S, 160°E–150°W) regions were constructed from the monthly Hadley Centre Sea Ice and Sea Surface Temperature (HadISST) dataset (Rayner et al. 2003).

Several different sets of NCEP Medium-Range Forecast (MRF9) and NCAR Community Climate Model (CCM3) AGCM integrations were also used (Table 1). A detailed description of the MRF9 model may be found in Kumar et al. (1996) and references therein, and of the CCM3 model in Kiehl et al. (1998).

One set of seasonal MRF9 integrations was made by Sardeshmukh et al. (2000) and Compo et al. (2001). These large ensembles were generated with prescribed observed monthly global climatological JFM SSTs (90 members) and observed monthly global SSTs for JFM 1987 and JFM 1989 (60 members each). A second set of 360 seasonal integrations, not previously reported, was generated with observed monthly global climatological JFM SSTs specified everywhere except in the Niño-4 region, where anomalies of $\pm 1°$, $\pm 3°$, and $\pm 5°$C were specified. Selected variables from both of these sets were archived at twice-daily resolution.

Four additional sets of Atmospheric Model Intercomparison Project (AMIP)-style AGCM integrations (i.e., runs with prescribed observed SSTs), available from two modeling centers at monthly resolution, were also used (Table 1). These include a 13-member MRF9 ensemble generated at NCEP for the period February 1950–February 1995 with observed monthly SSTs specified globally (commonly referred to as GOGA), and a similar nine-member ensemble with the observed monthly SSTs specified only in the tropical Pacific and climatological SSTs specified elsewhere (POGA). We also used a 12-member GOGA ensemble of CCM3 integrations made at NCAR for the period January 1950–December 1999, and a similar 11-member ensemble with the observed monthly SSTs specified only in the Tropics (30°N–30°S) and climatological SSTs specified elsewhere (TOGA).

To prepare these data for analysis, anomalies were calculated as departures from a least squares fit to the first three harmonics of the mean annual cycle over the 1950–79 period. For the AMIP-style integrations, the anomalies were computed separately for the POGA, TOGA, and two GOGA integrations as departures from their respective annual cycles. This base period was chosen to minimize the effects of the 1979 satellite discontinuity in the reanalysis dataset (Kistler et al. 2001), while utilizing the 30 years overlapping the AGCM datasets available for this study. Anomalies for the twice-daily MRF9 ensemble integrations were derived relative to the daily ensemble mean of the (90 member) climatological-SST integrations. All reanalysis and AGCM anomaly fields were then spatially smoothed to triangular truncation 31 using the spectral smoothing filter of Sardeshmukh and Hoskins (1984, hereafter SH).

Using the twice-daily data for each ensemble member (or each calendar year of the reanalysis), storm track values were defined at each grid point as the 2–7-day bandpass-filtered variance, and computed directly from the Fourier power spectrum of the 90-day JFM anomaly segments (Compo et al. 2001). These variance fields were then smoothed to triangular truncation 12 with the SH filter to facilitate comparison with other studies using similar truncations (e.g., Whitaker and Sardeshmukh 1998; Chang and Fu 2002, 2003). Our spatial smoothing retains about 70% of the original standard deviation, but the pattern is preserved. For example, the pattern correlation of the smoothed and unsmoothed reanalysis storm tracks is 0.98. In the following, the smoothed 2–7-day 500-mb vertical velocity variance fields are referred to as the omega storm track or simply as the storm track where there is no possibility of confusion.

All storm track anomaly fields were calculated after the spatial smoothing. Omega storm track anomalies for each JFM season in the reanalysis dataset were constructed by removing the 1950–79 JFM average storm track from each winter's storm track. For the MRF9 storm track anomalies, the JFM ensemble-mean storm track of the MRF9 climatological-SST ensemble was removed from each ensemble member's storm track.

Empirical orthogonal functions (EOFs) of the 570 MRF9 JFM storm track and 200-mb-height anomaly fields were computed separately for the Northern Hemisphere (20°–90°N), the Pacific–North American region (20°–90°N, 180°–60°W), and the North Atlantic–European region (20°–90°N, 60°W–60°E) using the covariance matrix area-weighted by the cosine of latitude. From the EOFs, the equivalent spatial degrees of freedom (esdof) for all three domains (relevant in assessing statistical significance) were calculated using the method of Bretherton et al. (1999).

To orient the reader to the climatological pattern and interannual variance of the omega storm tracks compared to the 500-mb-height storm tracks, the top half of Fig. 3 shows the 1950–79 averages of these measures of synoptic variability, and the bottom half shows their standard deviation over the period 1950–99. The mean omega storm track field in the top left is in excellent agreement with that determined by Hoskins and Hodges (2002) from the European Centre for Medium-Range Weather Forecasts (ECMWF) data for 1979–2000. This is consistent with Compo et al.'s earlier finding that the synoptic time-scale vertical velocity variance is similar between various observational estimates in the Northern Hemisphere extratropics. As illustrated by the shading in Fig. 3, the omega storm track field captures the well-known maxima in the Pacific and Atlantic, and also has a well-defined Mediterranean maximum seen in cyclone feature-tracking studies but not in climatologies of several other bandpass-filtered quantities such as 500-mb heights (Hoskins and Hodges 2002).

3. Empirical storm track model

a. Description of the empirical storm track model

Understanding the connection between a background flow and the behavior of individual synoptic eddies evolving on it has long been a core problem in dynamical meteorology. The shift of focus to the link between a mean flow and the general statistics of the synoptic eddies associated with it—storm tracks—is a relatively recent development (e.g., Blackmon et al. 1977; Lau 1988; Wallace et al. 1988; Farrell and Ioannou 1994, 1995; Branstator 1995; Whitaker and Sardeshmukh 1998; Zhang and Held 1999). Using a stochastically

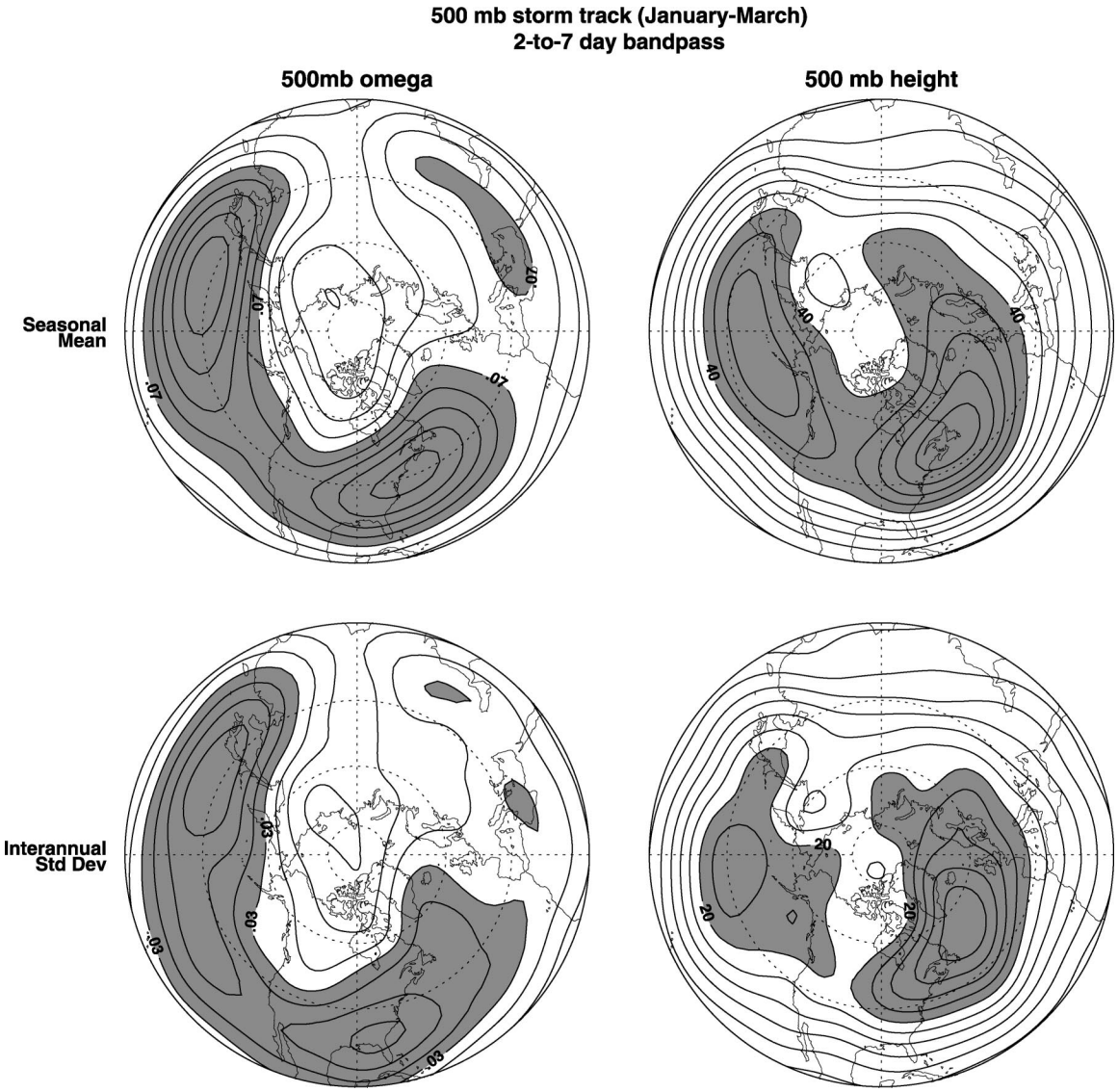

FIG. 3. (top) Climatological winter mean and (bottom) interannual standard deviation of 2–7-day bandpass variance of (left) 500-mb vertical velocity (omega) and (right) 500-mb height from NCEP–NCAR reanalyses. Note that the square root of each field is plotted. Contour intervals are (top left) 0.01 Pa s^{-1}, (top right) 5 m, (bottom left) 0.005 Pa s^{-1}, and (bottom right) 2.5 m. Shading highlights regions greater than (top left) 0.07 Pa s^{-1}, (top right) 40 m, (bottom left) 0.03 Pa s^{-1}, and (bottom right) 20 m.

forced two-layer quasigeostrophic model linearized about a specified climatological-mean flow, Whitaker and Sardeshmukh (1998) were able to simulate many aspects of the observed climatological Pacific and Atlantic storm tracks. Encouraged by this, they put their model to a harder test: to predict the anomalous storm tracks for individual winters given the anomalous winter-mean flow. Overall, they had only limited skill at this, raising the questions of whether this was due to the neglected nonlinearity of the mean-flow–storm track relationship, the relative simplicity of their model, the imprecise link between an *individual* winter's mean flow and storm tracks (i.e., intrinsically limited predictability), or some other factor.

Whitaker and Sardeshmukh used a dynamical storm track model linearized about a specified mean flow to deduce the storm tracks associated with that flow. One can also think of constructing an empirical storm track model that uses a multiple linear regression operator estimated from independent data to predict the anomalous storm tracks associated with an anomalous mean flow. The prediction equation may formally be written as

$$\mathbf{y} = \mathbf{G}\mathbf{x} + \varepsilon, \tag{2}$$

where $\mathbf{G}$ is the linear regression operator, $\mathbf{x}$ (the predictor) is the anomalous mean flow, $\mathbf{y}$ (the predictand) is the anomalous storm track field, and ε is the error.

Note that ε includes contributions from model truncation error and neglected nonlinearities, as well as sampling fluctuations; the latter can be made relatively small by ensemble and/or time averaging.

For the Northern Hemisphere winter (JFM), we have constructed such a model in a truncated EOF space, with **x** as the anomalous winter-mean 200-mb height and **y** as the anomalous winter-mean omega storm track, using the 570 NCEP AGCM integrations listed in Table 1. The optimal **G** was determined by cross-validation, sequentially removing 30 members of the set of 570 at a time, computing **G** from the remaining 540, and then predicting the storm track anomalies in the excluded 30. All EOF truncations from 2 to 60 for the 200-mb-height anomalies, and 2 to 70 for the storm track anomalies, were considered. The cross-validated root-mean-square error and average pattern correlation as a function of truncation are shown in Fig. 4. Although the cross-validated STM skill is not particularly sensitive to EOF truncation, we chose the truncation yielding the largest average pattern correlation: 40 EOFs of 200-mb heights and 51 EOFs of storm tracks. In EOF space, **G** is then a 40×51 matrix.

b. Validating the STM

To establish our linear STM's utility, we demonstrate its ability to deduce a nonlinear AGCM's SST-forced omega storm track signal given the AGCM's 200-mb height signal. Note that because our empirical STM is trained on an AGCM's *noise*, that is, on unpredictable storm track and mean-flow variations, its ability to predict SST-forced storm track variations is not guaranteed a priori, because 1) to the extent that the SST-forced mean-flow signal is weaker than the noise, one is putting the storm track model to an even harder test than the cross-validation test in Fig. 4; and more importantly, 2) the SST-forced 200-mb-height signal may have features not captured in the truncated EOF space in which **G** operates. Figure 5 shows that the STM is nevertheless able to predict the principal elements of the AGCM's storm track signal given the AGCM's 200-mb-height signal in (left) the El Niño winter of JFM 1987 and (right) the La Niña winter of JFM 1989. (To ensure that the verifications were performed on independent data, the EOFs and **G** for each of these two tests were rederived excluding the respective 60 members of the 1987 and 1989 AGCM integrations.) In both cases the hemispheric pattern correlation between the linear STM's diagnosed storm track anomalies and the nonlinear AGCM's ensemble-mean storm track anomalies is 0.9.

Our STM is complementary to that developed by Chang and Fu (2003), and also to that by Peng et al. (2003). Our approach is to construct the STM from much larger samples of independent AGCM statistics and use it to diagnose storm track anomalies from AGCM-simulated and observational 200-mb-height anomalies. Chang and Fu used a canonical correlation

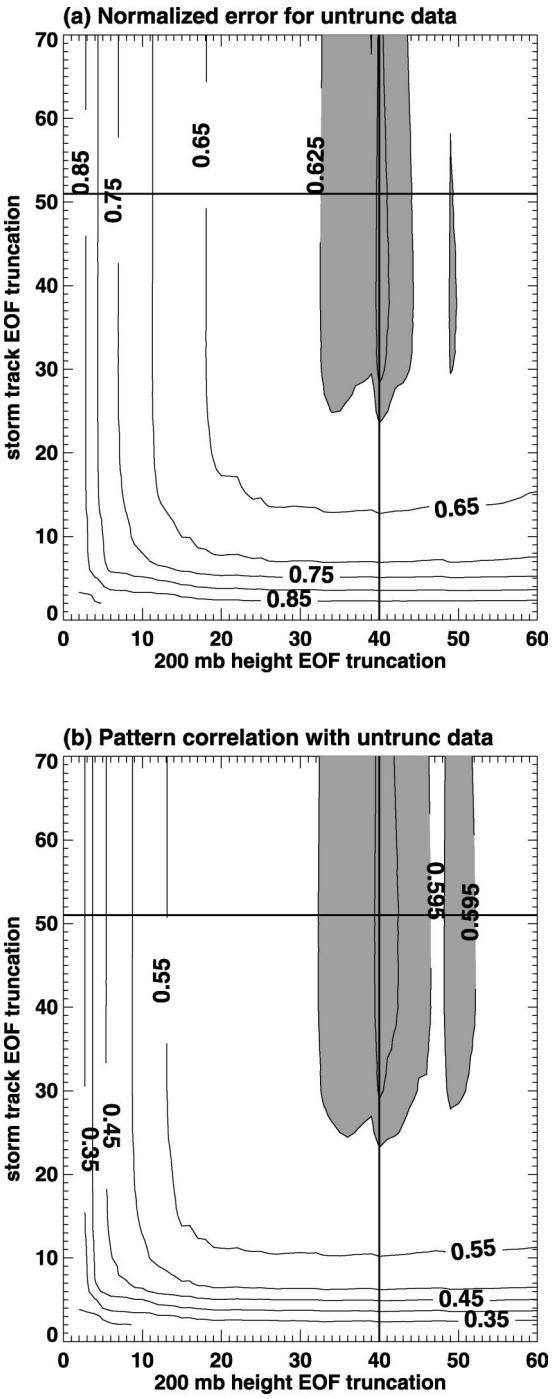

FIG. 4. Cross-validated skill of the empirical storm track model as a function of the number of retained EOFs of Jan–Mar 200-mb height seasonal mean and 500-mb vertical velocity bandpass variance seasonal anomalies. (a) Normalized error. (b) Anomaly pattern correlation. Contour interval is 0.05 with the addition of (a) 0.623, 0.625 and (b) 0.595, 0.598 contours. Shading begins at (a) 0.625 and (b) 0.595. Horizontal and vertical lines show the truncation used for the storm track model described in the text.

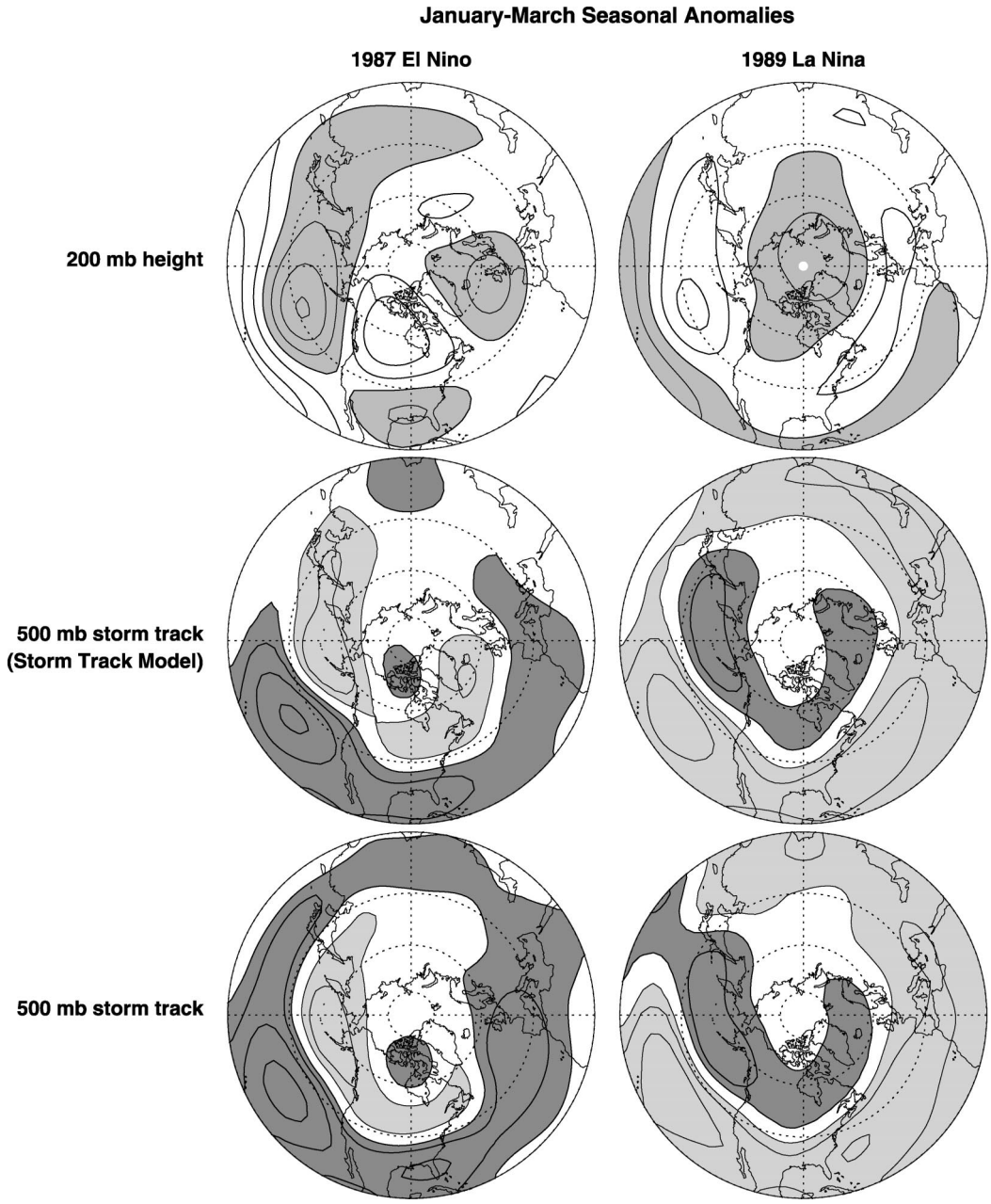

FIG. 5. Ensemble-mean seasonal anomalies for (left) Jan–Mar 1987 and (right) Jan–Mar 1989 AGCM ensemble of 60 members forced with 1987 and 1989 observed SSTs, respectively. (top) The 200-mb height anomaly; (middle) 500-mb storm track (vertical velocity 2–7-day bandpass variance anomaly) diagnosed using the empirical storm track model; (bottom) the 500-mb storm track. (middle and bottom) The plotted quantity is the signed square root of the variance anomaly. Contour intervals are (top) 20 m and (middle and bottom) 0.01 Pa s^{-1} with the zero contour suppressed. (top) Light shading indicates negative anomalies. (middle and bottom) Dark (light) shading indicates positive (negative) anomalies.

analysis (CCA) model of the anomalous mean flow and eddy statistics derived from a more recent part of the NCEP reanalysis dataset to assess the quality and decadal variability of eddy statistics in an earlier part. While successful in many respects, their model had substantial difficulty in capturing the amplitude of the observed storm track anomalies. Peng et al. (2003) used multiple linear regression on an NCEP AGCM dataset to construct a linear operator linking monthly mean geopotential heights and synoptic eddy vorticity fluxes to diagnose mean-flow/eddy feedbacks in that AGCM. Neither the storm track model of Peng et al. nor the

STM used here (Fig. 5) has the amplitude problem of Chang and Fu, which probably arises from sample size limitations of the observed record.

For the remainder of the paper, the STM used is that derived from all 570 members of the twice-daily MRF9 integrations (Table 1).

4. Skill in predicting interannual storm track variations

a. Local correlations

Having demonstrated that the STM can successfully diagnose an AGCM's ensemble-mean storm track anomalies in specific cases, we now use it to diagnose the Northern Hemisphere storm track anomalies in the past 50 winters from the observed winter-mean 200-mb-height anomalies, and also from the SST-forced 200-mb-height signals in each of those winters in four independent sets of AGCM integrations. The former will allow us to evaluate the average diagnostic skill of the STM in individual winters; the latter to estimate the predictability of interannual storm track variations, assuming perfect predictability of the SST variations.

Figure 6 shows, at each Northern Hemispheric grid point, the local temporal correlation of the observed and STM-diagnosed storm track anomalies. The contours begin at 0.25 (which is also the local 5% significance level) and are plotted every 0.15 thereafter. In Fig. 6 (top) the observed 200-mb-height anomalies are given to the STM to diagnose the storm track anomalies. In the other parts of Fig. 6, AGCM ensemble-mean 200-mb-height anomalies are given to the STM to estimate the AGCMs' ensemble-mean storm track anomalies. Figure 6 (top) shows that the STM's diagnostic skill is high over a large portion of the Northern Hemisphere, with correlations above 0.7 in regions of large storm track variability (Fig. 3). Its failure over the western Pacific and south Asia is consistent with a similar weakness in Fig. 5.

The bottom two rows of Fig. 6 show the actual skill in predicting the interannual storm track anomalies, using our STM-derived estimates of the AGCMs' ensemble-mean storm track anomalies as the prediction. Each panel shows the temporal correlation of the observed and predicted storm track anomalies. The correlations are very similar for the GOGA, TOGA, and POGA integrations, although a tendency for lower values in the latter two is apparent over the Gulf of Alaska and Europe. The results for the two AGCMs are generally consistent. The Pacific-only SST forcing reproduces almost all of the skill of the other integrations, suggesting that most of the storm track predictability arises from the tropical Pacific sector.

Figure 7 shows the expected skill ρ_{60} for the JFM 1987 and JFM 1989 storm track anomalies computed directly from Eq. (1) using the AGCM-predicted signal-to-noise ratio S from each 60-member ensemble. Comparing with the average skill over 50 winters in Fig. 6, it is evident that the expected forecast skill can be significantly different from the average skill and can also differ between El Niño and La Niña. Note, for example, that the expected skill over northern Europe is greater than 0.55 for the 1987 El Niño but less than 0.25 for the 1989 La Niña. (All differences of ~0.3 or more are significant above the 5% level, assuming S is distributed as a Student's t statistic.) Consistent with the altered atmospheric seasonal noise in response to ENSO found by Sardeshmukh et al. (2000), it is also interesting that the overall expected skill is higher in the La Niña case than the El Niño case over many regions not usually associated with a strong ENSO effect, such as the Atlantic, Africa, the Middle East, and south Asia. Considering only average skill masks such substantial—and potentially important—expected skill variations from case to case.

Comparing the expected skill maps in Fig. 7 for two particular forcing fields with any of the four AGCM actual average skill maps in Fig. 6 for 50 different forcing fields further illustrates the case dependence of skill. Given that Fig. 6 does not make the perfect model assumption, it is not surprising that all four actual skill maps have generally lower values than the expected skill maps in Fig. 7. What is perhaps more surprising is that in some regions (such as the Gulf of Mexico) the average actual skill in *all four maps* is actually *higher* than the expected skill in either map of Fig. 7. Clearly, once one goes beyond the broadbrush similarities of Figs. 6 and 7, substantial regional differences become apparent. They highlight the case-to-case dependence of storm track predictability, with important implications for the local precipitation predictability stressed previously (e.g., Fig. 2).

b. Time series of pattern correlation skill

To further characterize the interannual storm track variations, in Fig. 8 we examine the pattern correlation of the observed and predicted storm track anomalies for each winter in 1950–99 over the PNA and NATL-EUR sectors in Figs. 8c and 8d. In these panels the black, green, and blue bars indicate the pattern correlation between the observed storm track anomaly and that diagnosed from the STM using the observed, the ensemble-mean CCM3 GOGA, and the ensemble-mean CCM3 TOGA 200-mb-height anomalies, respectively. Results for the NCEP MRF9 (not shown) are similar. The high correlations obtained using the observed 200-mb-height field as the predictor show that the STM's diagnostic skill is substantial in both regions.

Figure 8 further supports the possibility of case-dependent storm track skill suggested in Fig. 7. The GOGA and TOGA integrations both simulate storm track anomalies with significant skill in many years over the PNA and NATL-EUR sectors. The case-to-case skill variations are large, and not entirely attributable to sam-

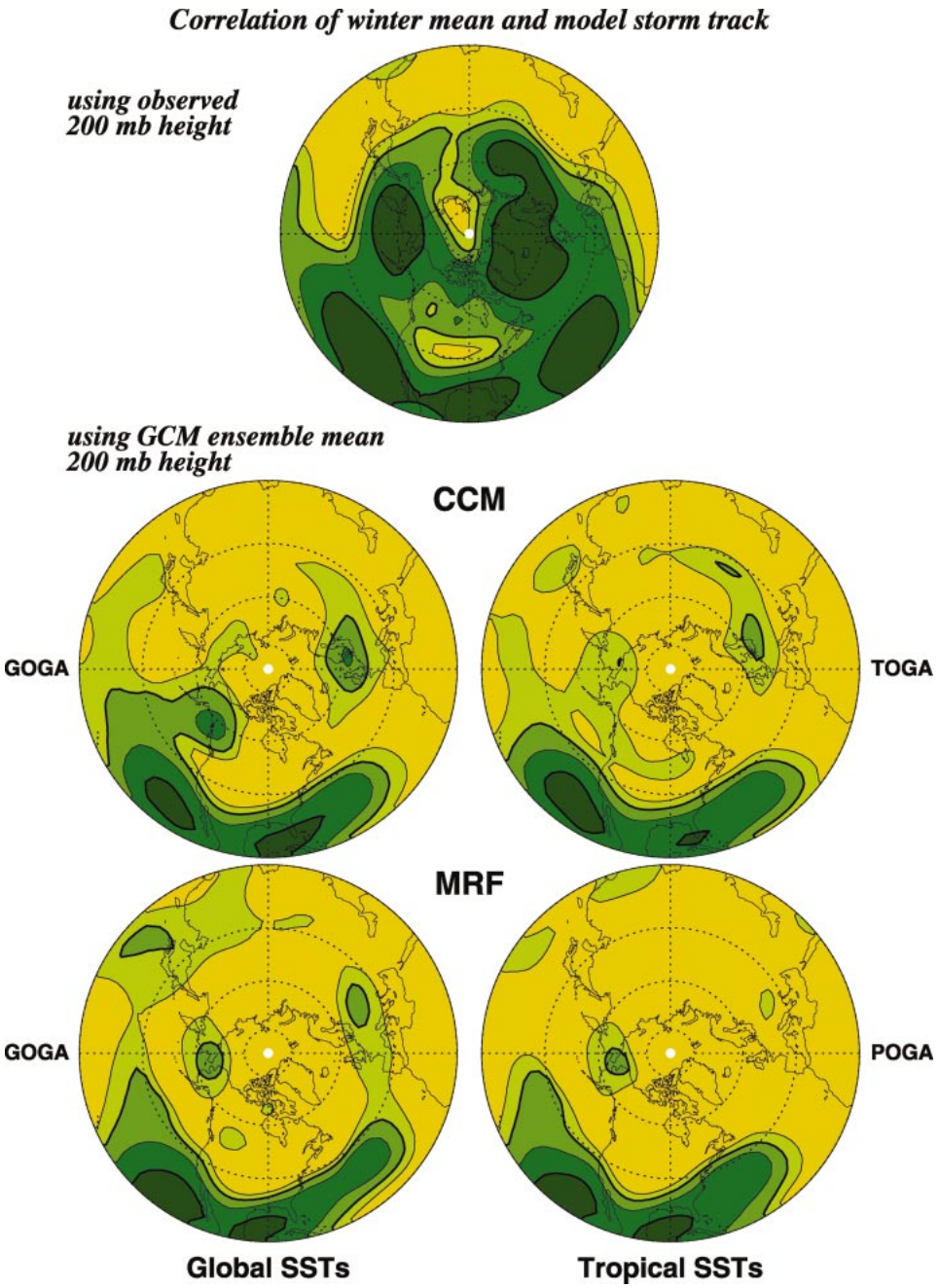

FIG. 6. Predictability of storm tracks estimated using the storm track model. Green shading begins at the 5% significance level of 0.25. Contour interval is 0.15 thereafter. The 0.4 and 0.7 contours are thickened for emphasis.

pling fluctuations. To establish this, we considered the null hypothesis that the simulated storm track anomalies are independent and bear no relationship to the observed anomalies. Conducting an extensive two-step Monte Carlo procedure using resampling with replacement, we first determined the distribution of skills arising from chance in 5000 Monte Carlo simulations of pattern correlations over the PNA and NATL-EUR sectors. For each realization in each sector, randomly selected GOGA and TOGA maps were correlated with the same randomly selected observed map to produce two random pattern correlations. From the distribution of pattern correlations, we found that both correlations *simultaneously* exceeded ~0.2 in only 5% of the 5000 simulations; this is indicated by the thin horizontal lines in Figs. 8c and 8d. We then performed a second and harder test [similar

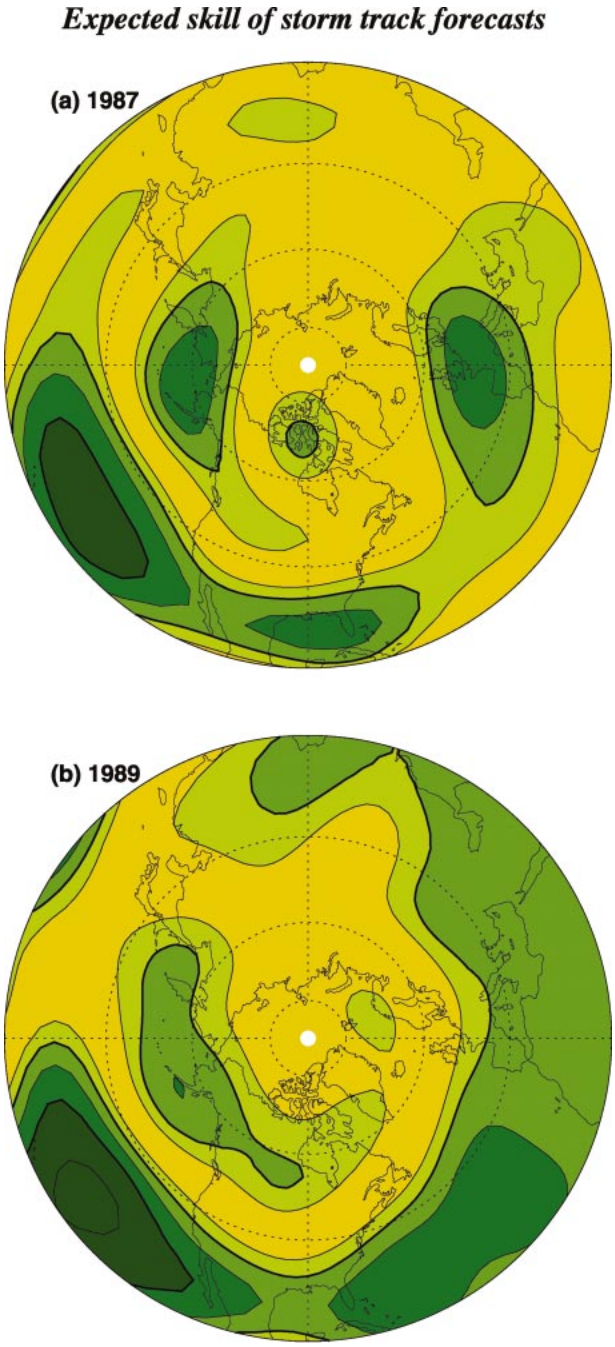

FIG. 7. Case dependence of the predictability of storm tracks. Comparison of expected local anomaly correlation skill for (a) Jan–Mar 1987 and (b) Jan–Mar 1989 storm track anomalies estimated using 60-member AGCM ensembles forced with observed SSTs to calculate the signal-to-noise ratio S and then applying Eq. (1) to calculate ρ_{60}. Contour interval is 0.15 starting at 0.25. The 0.4 and 0.7 contours are thickened for emphasis.

in spirit to the ''field significance'' test of Livezey and Chen (1983)], by determining the distribution of the number of times this ~0.2 correlation threshold was exceeded in 5000 simulations of 50-yr sets. We found that over the PNA and NATL-EUR sectors, 11 and 9 yr, respectively, out of 50 randomly passed the ~0.2 correlation threshold in 5% (250) of the 50-yr sets. In Fig. 8, 26 pairs of GOGA and TOGA integrations simultaneously exceed the correlation threshold in the PNA sector, and 12 pairs exceed the threshold in the NATL-EUR sector. The CCM3 skills in Fig. 8 over the PNA and NATL-EUR regions are thus significant above the 5% level.

Further investigation of these time series suggests that tropical Pacific SST variations are responsible at least partly for the storm track skill variations over the PNA and (to a lesser degree) the NATL-EUR sectors. Figure 8e shows the time series of winter-mean-standardized SST anomalies in the Niño-3 and Niño-4 regions. It is interesting that the PNA storm track anomalies are skillfully simulated in several years not usually classified as moderate to strong El Niño or La Niña events [e.g., National Oceanic and Atmospheric Administration (NOAA) Climate Prediction Center's subjective classification scheme at http://www.cpc.noaa.gov]. For example, in JFM 1970 and 1991 both the GOGA and TOGA runs have high skill, and yet both winters are classified as weak El Niño winters at CPC. Overall, the AGCM has significant skill in the PNA sector in 26 winters. This in itself is inconsistent with the idea that predictable signals occur only during moderate to strong ENSO events.

Barsugli and Sardeshmukh (2002) recently provided evidence that the extratropical circulation is relatively more sensitive to SST anomalies in Niño-4 than elsewhere in the tropical Indo-Pacific region. In particular, they found many aspects of the extratropical response to be nearly twice as sensitive to a Niño-4 anomaly as to a Niño-3 anomaly of the same magnitude. The dependence of the storm track skill in Fig. 8 on Niño-3 and Niño-4 anomalies is quantified in Table 2. The table shows, not surprisingly, that the largest average storm track skill in both the PNA and NATL-EUR sectors is obtained when both Niño-3 and Niño-4 anomalies are large. But interestingly, significant skill is also obtained in the PNA sector when the Niño-4 anomalies are large but the Niño-3 anomalies are not. These nine cases demonstrate that the predictable storm track signals do not arise solely from anomalous SST forcing in the classically defined regions of largest tropical SST variability (Niño-3 or Niño-3.4).

The results for the NATL-EUR sector in Table 2 are also interesting. As might perhaps have been anticipated from the top panel of Fig 6, the diagnostic skill is quite high throughout the record. The prediction skill, however, is generally low, consistent with some previous studies (e.g., Pavan and Doblas-Reyes 2000). Nonetheless, the unexpected moderate skill when both Niño-3 and Niño-4 anomalies are large suggests some sensitivity to tropical Pacific SST forcing.

Figures 7 and 8 and Table 2 suggest that storm track anomaly predictions will be more or less skillful de-

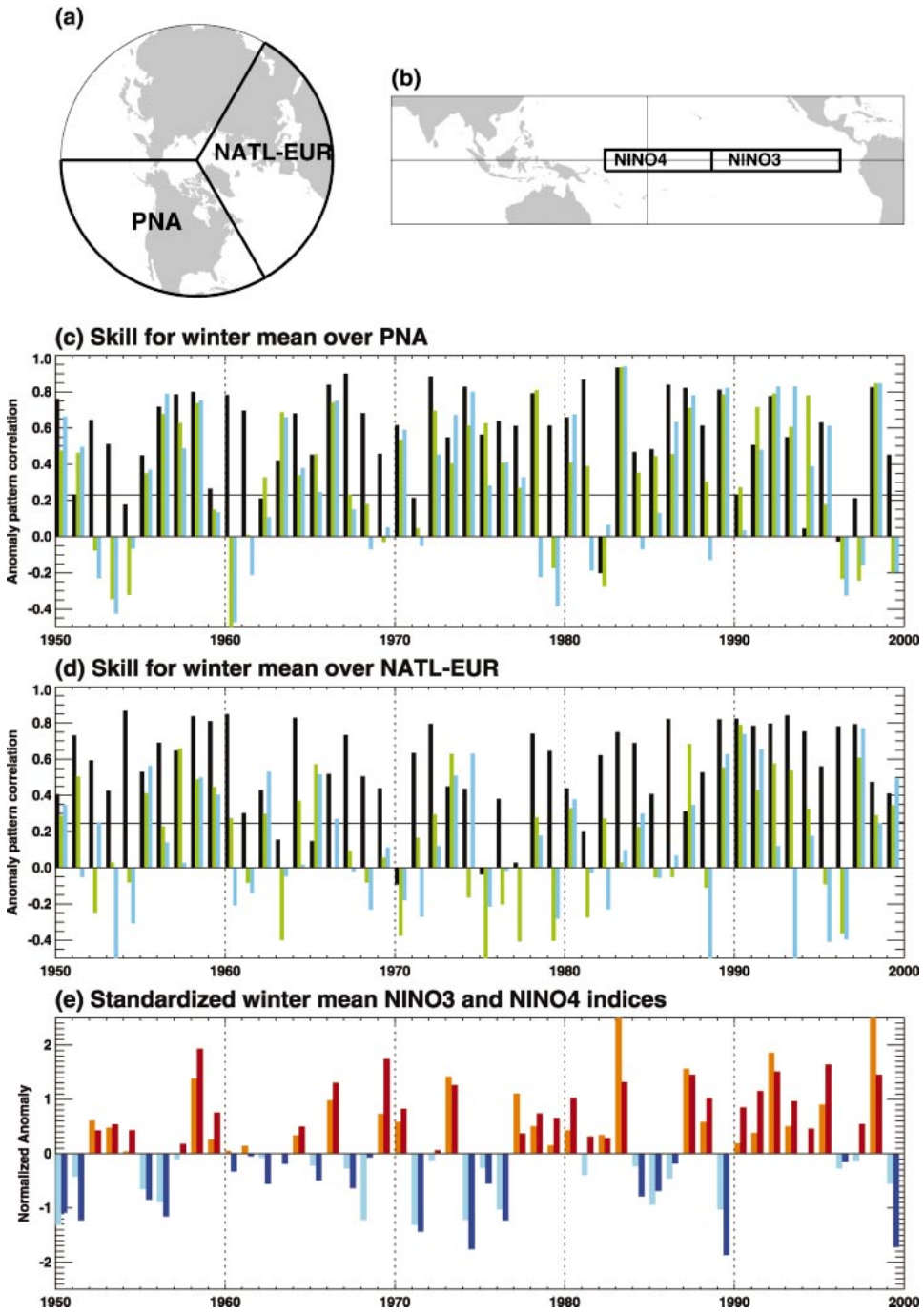

FIG. 8. Time series of anomaly pattern correlation for Jan–Mar (JFM) storm track anomalies. (a),(b) Regions used in subsequent panels. (c),(d) Pattern correlation over (c) the PNA and (d) NATL–EUR sectors between observed and STM-diagnosed storm track anomalies using 200-mb-height JFM anomalies from observed (black bars), CCM3 AGCM forced with global SSTs (green bars), CCM3 AGCM forced with tropical SSTs (blue bars). Thin horizontal line shows the 5% significance levels for both AGCM integrations having skill. (e) Time series of JFM Niño-3 (orange and light blue) and Niño-4 (red and dark blue) normalized by their respective standard deviations.

TABLE 2. Average storm track pattern correlation between the observed winter-mean storm track and that predicted by the STM given 200-mb-height anomaly fields from NCEP–NCAR reanalysis (OBS) and ensemble-mean anomalies of CCM3 tropical SST–forced (TOGA) and global SST–forced (GOGA) integrations from 1950–99. The skill is stratified by the magnitude of Niño-3 and Niño-4 indices and averaged separately over the PNA and NATL-EUR regions. Average correlations significant at or above the 5% level are indicated by bold italics.

	|Niño-3| $\geq 1\sigma$ |Niño-4| $\geq 1\sigma$	|Niño-3| $< 1\sigma$ |Niño-4| $\geq 1\sigma$	|Niño-3| $\geq 1\sigma$ |Niño-4| $< 1\sigma$	|Niño-3| $< 1\sigma$ |Niño-4| $< 1\sigma$
No. of cases	11	9	2	28
PNA				
TOGA	***0.68***	***0.39***	0.13	0.11
GOGA	***0.61***	***0.36***	0.23	***0.25***
OBS	***0.72***	***0.57***	***0.65***	***0.50***
NATL-EUR				
TOGA	***0.29***	0.11	−0.12	0.05
GOGA	***0.30***	0.19	−0.24	0.13
OBS	***0.57***	***0.57***	0.27	***0.57***

pending on the details of the SST forcing. To exploit this dependence in an operational setting, it will be important to anticipate such skill variations and issue them as part of the forecast. We will return to this point in section 6.

5. Skill in predicting decadal storm track variations

Perhaps we can push our empirical linear storm track model harder still, and ask it to diagnose decadal storm track variations (discussed, e.g., in Hurrell and van Loon 1997; Chang and Fu 2002) from the observed decadal mean flow changes. Chang and Fu (2003) had some success in this regard with their CCA storm track model derived from observed statistics, although the amplitude of the variations was weaker than observed. One would like to know how an STM such as ours derived from a much larger sample of AGCM statistics performs in this context. It is also important to quantify how much of the observed decadal storm track variations are linked to SST variations. These issues are addressed

The top section of Fig. 9 shows, similar to Fig. 6, the local temporal correlation of the observed five-winter-average storm track anomalies with those diagnosed by the STM given the observed anomalous five-winter-average 200-mb heights. Note that the contouring now begins at 0.4 (the local 5% significance level). The hemispheric coverage of significant five-winter-average diagnostic skill is much less than that for single winters. Nonetheless, most of the decadal variations over the North Atlantic and Europe are accurately diagnosed, with correlations over 0.85 in several regions. The STM also successfully diagnoses the five-winter-average variations over the eastern Pacific, but fails over large portions of western North America.

The bottom two rows of Fig. 9 show, again similar to Fig. 6, the actual skill in ''predicting'' the five-winter-mean storm track anomalies, using our STM-derived estimates of the AGCMs' ensemble-mean storm track anomalies as the ''prediction.'' Overall, these skill maps are very similar to each other and largely consistent with the interannual skill in Fig. 6, but with lower skill over the southern United States and higher skill over northern Canada. The TOGA SST forcing reproduces almost all of the skill of the other integrations, suggesting that most of the ''predictable'' decadal storm track signal arises from tropical SST forcing. A notable difference is the significant skill in the southern United States and western Atlantic in the CCM3 panels compared to the MRF9. This was also hinted in Fig. 6, and is perhaps related to the MRF9's known difficulty in reproducing observed upper-level features over the North Atlantic. A known slow mass leak at upper levels (Livezey et al. 1997) possibly also contributes to the lower MRF9 skill at this longer time scale.

The time series of pattern correlations of the observed and simulated five-winter-mean storm track anomalies over the PNA and NATL-EUR sectors are shown in Fig. 10, using the CCM3 GOGA and TOGA integrations as in Fig. 8. Also shown are the standardized five-winter-mean JFM values of the Niño-3 and Niño-4 SST anomalies. The GOGA and TOGA storm tracks have similar variations of skill over the PNA sector, strengthening the suggestion from Fig. 9 that the anomalous five-winter averages in much of the record are being forced by anomalous tropical SSTs. Comparing with the SSTs in Fig. 9 (bottom), periods of relatively high skill in the PNA sector correspond well with those of large Niño-3 and Niño-4 anomalies, leading one to suspect that the decadal variability of ENSO itself is driving most of this skill variation.

The STM's diagnostic skill has some noteworthy variations over the PNA sector, being consistently low in the mid-1960s and relatively high in the last 20 yr of the record. In contrast, except for a dip in the late 1960s, its diagnostic skill in the NATL-EUR sector is relatively high throughout the record, and generally much higher than the skill of the CCM3 integrations in the region.

While these variations of the diagnostic skill are interesting, it is difficult to ascertain if they are real, the result of STM deficiencies, inaccuracies in the observational 200-mb input heights, and/or in the verifying storm track output fields. It is easy to rationalize these sources of error: 1) The STM is trained on seasonal AGCM statistics whose EOFs may not adequately resolve decadal structures; 2) inhomogeneities in the reanalysis input data sources degrade the quality of the verifying synoptic variances (Chang and Fu 2003); and equally likely, 3) radiosonde discontinuities (e.g., Kistler et al. 2001; Harnik and Chang 2003), satellite discontinuities (Kistler et al. 2001), and changes in satellite retrieval methods over the period of record (Basist and

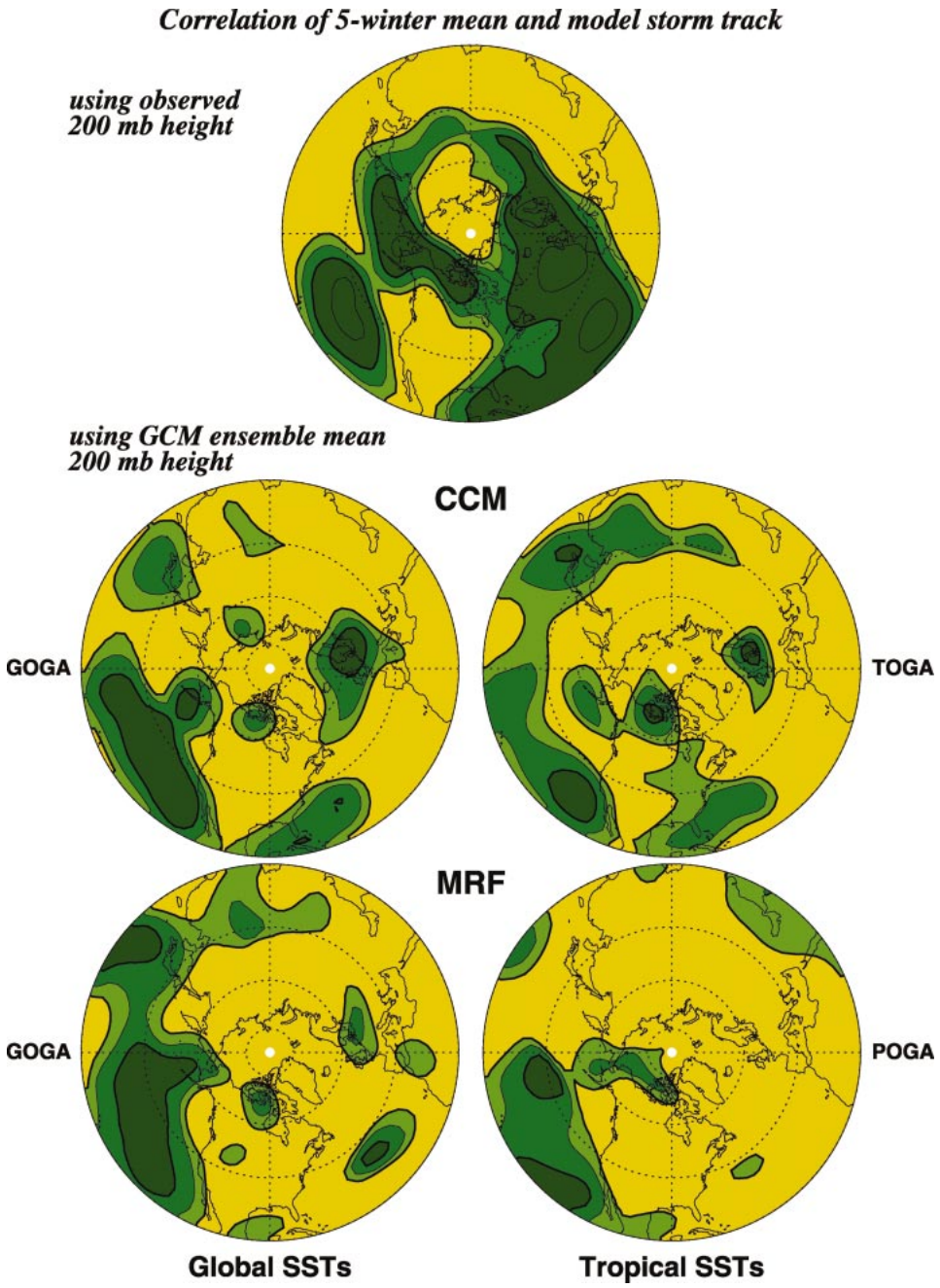

FIG. 9. Skill for five-winter averages. Green shading begins at the 5% significance level of 0.4. Contour interval is 0.15 thereafter.

Chelliah 1997) affect the quality of the decadal upper-tropospheric anomaly fields themselves.

Despite these concerns about the accuracy of the ''observed'' decadal storm track variations, there is evidence in Fig. 10 of the SST forcing of those variations over both the PNA and NATL-EUR sectors. However, whereas the skill is statistically significant over the record in the PNA sector (in the ''field significant'' sense discussed earlier), it is not so over the NATL-EUR sector. To determine the significance, we performed, as for Fig. 8, a Monte Carlo procedure in each sector to test the null hypothesis that the five-winter averages of both the GOGA and TOGA integrations are independent and have no relationship to the observed anomalies. From 5000 resamplings with replacement of the observed and predicted five-winter-mean storm tracks, we found that only 5% of both integrations simultaneously exceeded pattern correlations of ~0.3, shown by the thin hori-

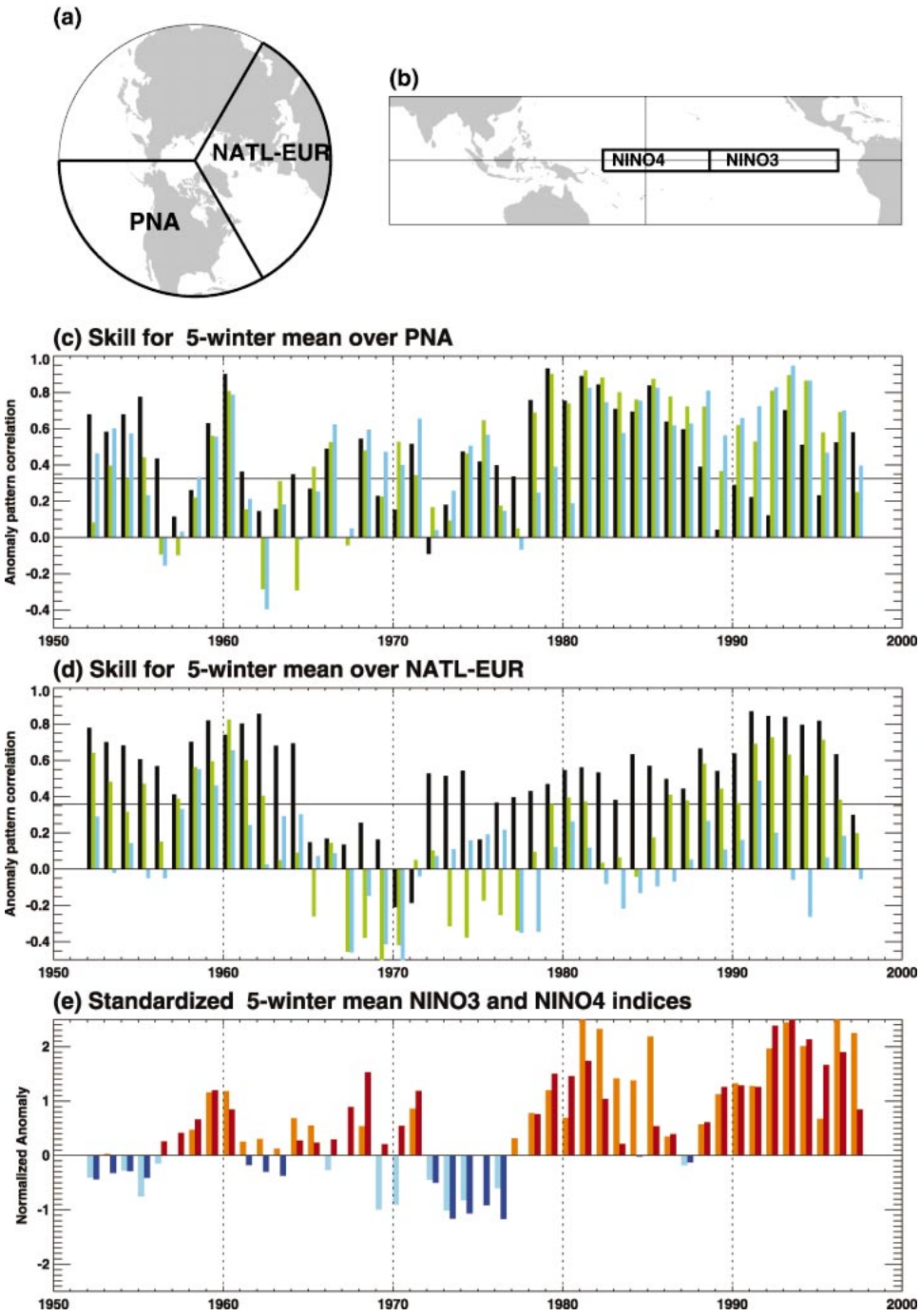

FIG. 10. As in Fig. 8, but for five-winter averages.

zontal lines in Figs. 10c and 10d. In a 46-yr sample of five-winter averages, an additional 5000 resamplings showed that 14 and 10 yr would randomly pass the threshold more than 5% of the time in the PNA and NATL-EUR sectors, respectively. The GOGA and TOGA skills in Fig. 10 simultaneously exceed the threshold 27 times over the PNA sector, but only 4 times over the NATL-EUR sector.

6. Reliability of storm track predictability estimates

The demonstration of predictable SST-forced storm track signals on interannual scales is a central result of this paper. The general similarity of the pattern of average skill in the AGCM parts of Fig. 6 with that of the expected (i.e., the potential) skill in two individual cases

in Fig. 7 is reassuring. But how realistic and representative are the expected skills in Fig. 7? The issue is important, since the gap between actual and potential skill sets targets for AGCM improvement. At the same time, it is important to recognize that all estimates of potential skill are ultimately model dependent, since they assume that the particular imperfect model used to make them is ''perfect.'' As discussed earlier in Fig. 1, they can also be compromised by using inadequate ensemble sizes, although this is much less of a concern in Fig. 7.

One way to address this issue would be to repeat the 60-member ensemble integrations of Fig. 7 for all the 50+ winters in the record with a large number of different AGCMs. This would make it possible to construct an average potential skill plot in which one might have more faith. However, errors common to the models would still compromise such a plot.

Here we pursue an alternative strategy to assess the reliability of our predictability estimates. In principle, this strategy should be applicable even to estimates generated using a single AGCM. The basic idea is to assess the degree to which the variations of actual skill are consistent with the variations of expected skill, that is, with variations of the signal-to-noise ratio *S*. Given *S* and the actual skill ρ for each winter, one could examine the extent to which a scatterplot of ρ against *S* follows the ρ_∞ curve (or more appropriately, the ρ_{60} curve if using 60-member ensembles) in Fig. 1. To that extent, our predictions would be deemed ''reliable'' (in the technical sense often used in probabilistic forecasting), and so would our predictability estimates. Additionally, if the variations of *S* from case to case are substantial, (i.e., if our forecasts have ''resolution'' in the technical sense), one would also have some skill in predicting the *variations* of skill from case to case, and specifically, in identifying highly skillful predictions a priori. If, on the other hand, the scatter points fall well below the ρ_{60} curve, one would have to conclude that model error was compromising our predictability estimates.

To our knowledge it is not yet possible to make such scatterplots, because the necessary large ensemble integrations with archived daily output have not yet been made at modeling centers. We have nevertheless attempted with Fig. 11 to generate such a plot using the smaller ~12 member ensemble runs of the NCEP and NCAR AGCMs for 1950–99 available to us. This figure provides perhaps the best overall assessment to date of Northern Hemispheric storm track predictability, although it will be clear from what follows that much else remains to be done.

The thick curve in Fig. 11 shows the expected skill for a 12-member ensemble as a function of *S*, assuming a perfect model. It is also possible to calculate the expected skill of a model with a time-varying systematic error [i.e., an error in representing the correct ensemble mean in each forecast case; see Sardeshmukh et al. (2000)]. As discussed in the appendix, the expected skill in this scenario is

$$\rho_n = S^2/[(S^2 + 1)(S^2 + S_e^2 + n^{-1})]^{1/2}, \qquad (3)$$

where S_e is the ratio of the ensemble-mean error to the ensemble spread. The dotted curve illustrates the expected skill for a 12-member ensemble if $S_e = 2S$.

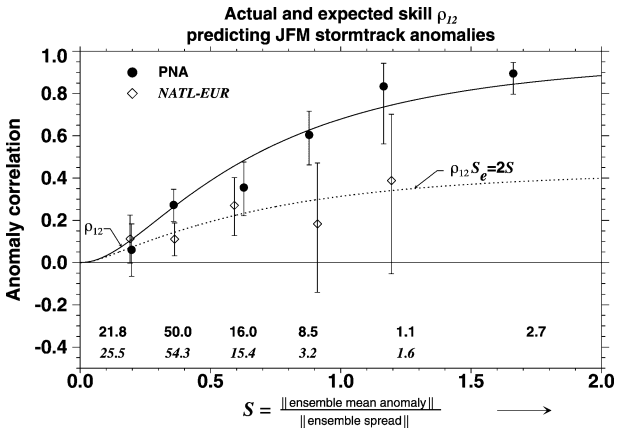

FIG. 11. Pattern anomaly correlation skill of storm track forecasts made using the CCM3 and MRF9 diagnosed storm tracks for the Jan–Mar season. Solid curve shows the expected correlation skill ρ_n of forecasts made from the mean of $n = 12$ member ensembles as a function of the signal-to-noise ratio *S* based on Eq. (1). Dotted curve shows the expected skill ρ_{12} when a systematic error $S_e = 2S$ is present in the forecast based on Eq. (3). Symbols show the actual skill of storm track forecasts for the PNA region (filled circles) and NATL-EUR region (diamonds) binned over similar *S* values. Bin widths are 0.25 from $S = 0$ to $S = 1$ and 0.5 thereafter. Percentage of cases in each bin is indicated. Error bars show the 95% confidence interval using the Fisher *z* transformation and assuming 6 esdof.

The plotted symbols in Fig. 11 show the average actual simulation skill of the storm track anomalies in the four AMIP-style integrations. They represent averages over neighboring *S* values in the PNA sector (filled circles) and the NATL-EUR sector (diamonds). To calculate *S*, the signal in each winter was taken as the STM-diagnosed ensemble-mean storm track anomaly. Unfortunately, the ensemble size of ~12 was too small for us to accurately deduce the storm track ensemble spread from the spread of the 200-mb heights. To estimate the noise, we therefore averaged the ensemble spreads in our 60-member 1987, 60-member 1989, and 90-member climatological SST MRF9 ensembles (Table 1), and used this as a constant spread in all the *S* calculations.

The most important result in Fig. 11 is that the actual storm track skill is generally consistent with the estimated *S* values in the PNA sector, but not in the NATL–EUR sector. A similar result is obtained for 5-yr averages, but with larger error bars (not shown). These results are insensitive to the inclusion or exclusion of the tropically forced runs (not shown). They also hold when the MRF9 and CCM3 runs are considered separately (not shown).

The magnitude of the ensemble-mean error over the

Atlantic sector appears to be about twice that of the AGCMs' predicted signals. Because the storm track signals are not obtained directly from the AGCMs but are diagnosed from the AGCMs' ensemble-mean 200-mb heights using the STM, the error lies either in the STM or the AGCMs' 200-mb heights. We believe an error in the latter is more likely, for two reasons: 1) The STM's diagnostic skill is quite high over the Atlantic sector (Figs. 6 and 8), where it is about the same on average as over the PNA sector (Table 2), and 2) the STM also successfully recovers the MRF9 anomalous ensemble-mean Atlantic storm track given the anomalous ensemble-mean 200-mb heights in the 1987 and 1989 cases (Fig. 5). These results suggest that both the MRF9 and CCM3 may have substantial errors in their 200-mb-height responses over the NATL-EUR region to specified global and tropical SSTs.

Many AGCMs have limited skill in predicting upper-level circulation and precipitation anomalies over the North Atlantic and Europe (e.g., Livezey et al. 1997; Brankovic and Palmer 2000; Doblas-Reyes et al. 2000; Graham et al. 2000; Peng et al. 2000; Shukla et al. 2000). The estimated S values in Fig. 11, as well as the expected skill in Fig. 7, suggest that if this model error were to be eliminated, the actual skill in this region would be higher than found here for storm track anomalies, and possibly for other related quantities such as winter-mean heights and precipitation as well. Note again that the AGCM error diagnosed in Fig. 11 is not a constant bias that can be linearly removed but varies from forecast case to case, which complicates its diagnosis and elimination.

7. Conclusions

Our study shows that there is a predictable SST-forced storm track signal over much of the Northern Hemisphere in boreal winter, but whose magnitude and pattern may differ substantially from case to case and between El Niño and La Niña events. The signal is modest on average, but has large implications for the prediction of seasonal precipitation. Our results further suggest that some predictable storm track signal may exist over the PNA sector even in weak and non-ENSO winters. These predictable signals are largely associated with predictable anomalous tropical SSTs, with a minor but statistically significant contribution by anomalous extratropical SSTs. The case-to-case variations of the signal are sufficiently large to be useful in anticipating which storm track forecasts are likely to be more (or less) skillful than on average. We have also presented quantitative evidence that the decadal storm track variations in the PNA sector are consistent with tropical SST variations, particularly in the second half of the record. To what degree those SST variations are themselves predictable is unknown at present.

We have not been able to establish these conclusions in the North Atlantic–European sector. While a potentially predictable storm track signal may exist there, a substantial systematic error, on order twice the signal strength, has been diagnosed in both the AGCMs used in this study. If the error is eliminated, then judging by the potential skill results in Figs. 7 and 11, the storm track skill in this region could improve well beyond the level shown here.

Our predictability study is founded on two relationships. The first is Eq. (1), which relates the signal-to-noise ratio S to the expected correlation skill ρ_n for any forecast variable at any grid point, or any combinations of variables at any combinations of grid points, *including quadratic measures of bandpass-filtered quantities such as storm tracks.* The value S for the storm tracks then serves as a useful simple measure of storm track predictability.

The second foundation of this study, consistent with the success of recent empirical STMs (Chang and Fu 2003) and our own calculations, is that the relationship between anomalous mean flows and anomalous storm tracks, expressed in Eq. (2), is essentially a linear one in a relatively low-dimensional space. It is remarkable how well our linear STM reproduces a nonlinear AGCM's storm track response to ENSO forcing given only the AGCM's 200-mb-height response. It is interesting that this simple empirical STM is also significantly better at replicating observed anomalous storm tracks in individual winters than the dynamical STM of Whitaker and Sardeshmukh (1998). This should prove useful for diagnosing the errors of Whitaker and Sardeshmukh's STM and other dynamical STMs.

We have relied heavily on these two relationships to estimate the local and regional predictability of winter-mean and five-winter-mean storm track anomalies. Perhaps most importantly, they have enabled us to construct Fig. 11, with its large implications for the potential for improvement in AGCM storm track (and precipitation) prediction skill over the North Atlantic and Europe. The results in Fig. 11 for the PNA sector increase our confidence in the reliability of our predictability estimates in that sector. Figure 11 further suggests that at least some of the time variation of actual skill in Fig. 8 may itself be predictable via a priori estimation of S. As discussed in the appendix, an ensemble size of order 128 members should be adequate to accurately estimate winter-to-winter variations of S as small as 0.25, and therefore variations of expected skill of a similar magnitude.

Finally, it should be recognized that S is between 0.25 and 0.75 in about 65% of the 50+ winters considered in Fig. 11. As the theoretical curves in Fig. 11 (and Fig. 1) show, when S is small the skill of deterministic ensemble-mean predictions is low, and is further compromised by model error and the use of small ensembles. With large ensembles and small model error, useful *probabilistic* predictions, especially of the altered risks of extreme anomalies, are however still possible, as stressed by Sardeshmukh et al. (2000). It would there-

fore be worthwhile to improve the actual storm track prediction skill even for the low S situations in Fig. 11.

Acknowledgments. The authors would like to acknowledge useful discussions with colleagues at CDC, particularly M. Alexander, J. Barsugli, J. Bergman, M. Newman, C. Penland, and J. Whitaker. The comments of three anonymous reviewers and the editor, S. Schubert, greatly improved the clarity of the manuscript. D. Hooper, C. Smith, and G. Bates provided invaluable assistance with data processing. The HadISST data were kindly provided by N. Rayner and the Hadley Centre of the Met Office. NCEP–NCAR reanalysis data were obtained from the CDC archives (http://www.cdc.noaa.gov). NCEP MRF9 AMIP runs were made courtesy of M. Ji of NCEP. NCAR CCM3.0 runs were provided courtesy of Y.-H. Lee of NCAR. This work was partially supported by a grant from NOAA's CLIVAR-Pacific Program.

APPENDIX

Derivation of Multivariate Predictability as a Function of the Signal-to-Noise Ratio

One way to increase the value of climate predictions is to issue the expected skill of the forecast as part of the prediction. Van den Dool and Toth (1991) derived the expected value of a forecast's correlation skill when predicting the ensemble-mean anomaly of any univariate distribution using an infinite-member ensemble. Rowell (1998) extended Van den Dool and Toth's result to the use of an n-member ensemble forecast. Kumar and Hoerling (2000) derived the results of Rowell (1998) for the specific case of forecasting the sign of the mean anomaly of a Gaussian distribution.

Sardeshmukh et al. (2000) further developed the results of Rowell (1998) for a multivariate forecast with errors. They derived the expected skill for an ensemble forecasting system predicting any multivariate quantity that has defined first and second moments. The forecast model need not be perfect. Following Sardeshmukh et al., consider a multivariate distribution $P_m(\langle \mathbf{x} + \mathbf{x}_e \rangle, \mathbf{C}_{0m})$ that represents the altered model probability density function (PDF) of some quantity $\mathbf{x}$ and $P(\langle \mathbf{x} \rangle, \mathbf{C}_0)$ that is the true PDF of that quantity, such as winter storm track anomalies during an El Niño event. Here $\langle \mathbf{x} \rangle$ is the population-mean anomaly-state vector, $\langle \mathbf{x} + \mathbf{x}_e \rangle$ is the model's population-mean anomaly-state vector, $\langle \mathbf{x}_e \rangle$ is the model's error in predicting the population mean, $\mathbf{C}_0$ is the covariance matrix of the variations $\mathbf{x}'$ around $\langle \mathbf{x} \rangle$, and $\mathbf{C}_{0m}$ is the model's covariance matrix of variations $\mathbf{y}'$ about $\langle \mathbf{x} + \mathbf{x}_e \rangle$. Note that P and P_m can be any multivariate distributions with defined first and second moments. Also, note that none of the parameters of these distributions need necessarily be the same for El Niño and La Niña or even from case to case. The PDF of ensemble-mean forecasts issued from an n-member ensemble with this model is $P_m(\langle \mathbf{x} + \mathbf{x}_e \rangle, n^{-1}\,\mathbf{C}_{0m})$. Assume that a vector $\mathbf{y}$ is issued as the ensemble-mean forecast, and the real atmosphere picks a vector $\mathbf{x} = \langle \mathbf{x} \rangle + \mathbf{x}'$ from P as its storm track anomaly field. The average anomaly correlation of the observed and predicted vectors is then

$$\rho_n = \frac{\langle \mathbf{x} \cdot \mathbf{y} \rangle}{(\langle \mathbf{x} \cdot \mathbf{x} \rangle \langle \mathbf{y} \cdot \mathbf{y} \rangle)^{1/2}} = \frac{\langle \mathbf{x} \rangle \cdot \langle \mathbf{x} \rangle}{[(\langle \mathbf{x} \rangle \cdot \langle \mathbf{x} \rangle + \langle \mathbf{x}' \cdot \mathbf{x}' \rangle)(\langle \mathbf{x} \rangle \cdot \langle \mathbf{x} \rangle + \langle \mathbf{x}_e \rangle \cdot \langle \mathbf{x}_e \rangle + n^{-1} \langle \mathbf{y}' \cdot \mathbf{y}' \rangle)]^{1/2}}, \tag{A1}$$

where we have assumed that $\langle \mathbf{x} \rangle \cdot \langle \mathbf{x}_e \rangle = 0$. The dot product here represents a general scalar product of the form $\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^{\mathrm{T}} \mathbf{W} \mathbf{y}$, where $\mathbf{W}$ is any suitable positive-definite weight matrix. Note that $\langle \mathbf{x}' \cdot \mathbf{x}' \rangle = \mathrm{Tr}[\mathbf{W}^{1/2} \mathbf{C}_0 \mathbf{W}^{1/2}]$ and $\langle \mathbf{y}' \cdot \mathbf{y}' \rangle = \mathrm{Tr}[\mathbf{W}^{1/2} \mathbf{C}_{0m} \mathbf{W}^{1/2}]$. The weight matrix $\mathbf{W}$ can be chosen to emphasize a particular grid point (as in Fig. 7), a linear combination of variables over a region (as in Fig. 11), or be set equal to identity to examine skill over the entire atmosphere.

If we further assume that the model correctly reproduces the second moment, that is, $\mathbf{C}_{0m} = \mathbf{C}_0$, then (A1) becomes (3):

$$\rho_n = S^2/[(S^2 + 1)(S^2 + S_e^2 + n^{-1})]^{1/2}, \tag{A2}$$

where $S = [\langle \mathbf{x} \rangle \cdot \langle \mathbf{x} \rangle / \langle \mathbf{x}' \cdot \mathbf{x}' \rangle]^{1/2}$ and $S_e = [\langle \mathbf{x}_e \rangle \cdot \langle \mathbf{x}_e \rangle / \langle \mathbf{x}' \cdot \mathbf{x}' \rangle]^{1/2}$. For a "perfect" model, both $\langle \mathbf{x}_e \rangle = 0$ and $\mathbf{C}_{0m} = \mathbf{C}_0$, and (A1) leads to (1):

$$\rho_n = S^2/[(S^2 + 1)(S^2 + n^{-1})]^{1/2}. \tag{A3}$$

In the limit as the ensemble size goes to infinity, for a perfect model $\rho_\infty = S/\sqrt{(S^2 + 1)}$. For univariate distributions, ρ_∞^2 is closely related to the predictability measure examined by Koster et al. (2000).

Graham et al. (2000) found that increasing the number of ensemble members did little to increase skill, even in the context of a perfect model. This can be understood directly in terms of ρ_n in (A3) and graphically as illustrated in Fig. 1. Graham et al. (2000) empirically calculated ρ_n with $n = 9$ and $n = 18$ and compared it to their actual anomaly correlation (their Figs. 13 and 14). It is clear from (A3) and Fig. 1 that little is expected to be gained by increasing ensemble size from 9 to 18 members. As reviewed in Sardeshmukh et al. (2000), if the signal is considered the difference of two means each of size n, then S is distributed as a Student t statistic, and the 95% confidence interval about S is approximately $\pm 2(2/n)^{1/2}$. The advantage, then, of using much larger ($n \geq 128$) ensemble sizes lies in improving

the ability to forecast forecast skill by reducing the uncertainty in S illustrated in Fig. 1. Note also from (A2) that any advantage for actual skill can be lost from model error, as illustrated by the dotted curve in Fig. 11.

REFERENCES

Barsugli, J. J., and P. D. Sardeshmukh, 2002: Global atmospheric sensitivity to tropical SST anomalies throughout the Indo-Pacific basin. *J. Climate,* **15,** 3427–3442.

Basist, A. N., and M. Chelliah, 1997: Comparison of tropospheric temperatures derived from the NCEP/NCAR reanalysis, NCEP operational analysis, and the Microwave Sounding Unit. *Bull. Amer. Meteor. Soc.,* **78,** 1431–1447.

Blackmon, M. L., J. M. Wallace, N.-C. Lau, and S. L. Mullen, 1977: An observational study of the Northern Hemisphere wintertime circulation. *J. Atmos. Sci.,* **34,** 1040–1053.

Brankovic, C., and T. N. Palmer, 2000: Seasonal skill and predictability of ECMWF PROVOST ensembles. *Quart. J. Roy. Meteor. Soc.,* **126,** 2035–2067.

Branstator, G., 1995: Organization of storm track anomalies by recurring low-frequency circulation anomalies. *J. Atmos. Sci.,* **52,** 207–226.

Bretherton, C. S., M. Widmann, V. P. Dymnikov, J. M. Wallace, and I. Blade, 1999: The effective number of spatial degrees of freedom of a time-varying field. *J. Climate,* **12,** 1990–2009.

Carillo, A., P. M. Ruti, and A. Navarra, 2000: Storm tracks and zonal mean flow variability: A comparison between observed and simulated data. *Climate Dyn.,* **16,** 219–228.

Chang, E. K. M., and Y. Fu, 2002: Interdecadal variations in Northern Hemisphere winter storm track intensity. *J. Climate,* **15,** 642–658.

——, and ——, 2003: Using mean flow change as a proxy to infer interdecadal storm track variability. *J. Climate,* **16,** 2178–2196.

Compo, G. P., P. D. Sardeshmukh, and C. Penland, 2001: Changes of subseasonal variability associated with El Niño. *J. Climate,* **14,** 3356–3374.

Doblas-Reyes, F. J., M. Deque, and J.-P. Piedelievre, 2000: Multi-model spread and probabilistic seasonal forecasts in PROVOST. *Quart. J. Roy. Meteor. Soc.,* **126,** 2069–2087.

Farrell, B. F., and P. J. Ioannou, 1994: A theory for the statistical equilibrium energy spectrum and heat flux produced by transient baroclinic waves. *J. Atmos. Sci.,* **51,** 2685–2698.

——, and ——, 1995: Stochastic dynamics of the midlatitude atmospheric jet. *J. Atmos. Sci.,* **52,** 1642–1656.

Fraedrich, K., 1994: An ENSO impact on Europe? A review. *Tellus,* **46A,** 541–552.

——, and K. Muller, 1992: Climate anomalies in Europe associated with ENSO extremes. *Int. J. Climatol.,* **12,** 25–31.

Graham, N. E., and H. F. Diaz, 2001: Evidence for intensification of North Pacific winter cyclones since 1948. *Bull. Amer. Meteor. Soc.,* **82,** 1869–1893.

Graham, R. J., A. D. L. Evans, K. R. Mylne, M. S. J. Harrison, and K. B. Robertson, 2000: An assessment of seasonal predictability using atmospheric general circulation models. *Quart. J. Roy. Meteor. Soc.,* **126,** 2211–2240.

Gulev, S. K., T. Jung, and E. Ruprecht, 2002: Climatology and interannual variability in the intensity of synoptic-scale processes in the North Atlantic from the NCEP–NCAR reanalysis data. *J. Climate,* **15,** 809–828.

Harnik, N., and E. K. M. Chang, 2003: Storm track variations as seen in radiosonde observations and reanalysis data. *J. Climate,* **16,** 480–495.

Hoerling, M. P., and M. Ting, 1994: Organization of extratropical transients during El Niño. *J. Climate,* **7,** 745–766.

Hoskins, B. J., and K. I. Hodges, 2002: New perspectives on the Northern Hemisphere winter storm tracks. *J. Atmos. Sci.,* **59,** 1041–1061.

Hurrell, J. W., and H. van Loon, 1997: Decadal variations of climate associated with the North Atlantic Oscillation. *Climatic Change,* **36,** 301–326.

Kiehl, J. T., J. J. Hack, G. B. Bonan, B. A. Boville, D. L. Williamson, and P. J. Rasch, 1998: The National Center for Atmospheric Research Community Climate Model: CCM3. *J. Climate,* **11,** 1131–1150.

Kistler, R., and Coauthors, 2001: The NCEP–NCAR 50-year reanalysis: Monthly means CD-ROM and documentation. *Bull. Amer. Meteor. Soc.,* **82,** 247–267.

Koster, R. D., M. J. Suarez, and M. Heiser, 2000: Variance and predictability of precipitation at seasonal-to-interannual timescales. *J. Hydrometeor.,* **1,** 26–46.

Kumar, A., and M. P. Hoerling, 1998: Specification of regional sea surface temperatures in atmospheric general circulation model simulations. *J. Geophys. Res.,* **103,** 8901–8907.

——, and ——, 2000: Analysis of a conceptual model of seasonal climate variability and implications for seasonal prediction. *Bull. Amer. Meteor. Soc.,* **81,** 255–264.

——, ——, M. Ji, A. Leetma, and P. Sardeshmukh, 1996: Assessing a GCM's suitability for making seasonal predictions. *J. Climate,* **9,** 115–129.

Lau, N. C., 1988: Variability of the observed midlatitude storm tracks in relation to low-frequency changes in the circulation pattern. *J. Atmos. Sci.,* **45,** 2718–2743.

Livezey, R. E., and W. Y. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.,* **111,** 46–59.

——, M. Masutani, A. Leetmaa, H. Rui, M. Ji, and A. Kumar, 1997: Teleconnective response of the Pacific–North American region atmosphere to large central equatorial Pacific SST anomalies. *J. Climate,* **10,** 1787–1820.

Matthews, A. J., and G. N. Kiladis, 1999: Interaction between ENSO, transient circulation, and tropical convection over the Pacific. *J. Climate,* **12,** 3062–3086.

May, W., and L. Bengtsson, 1998: The signature of ENSO in the Northern Hemisphere midlatitude seasonal mean flow and high-frequency intraseasonal variability. *Meteor. Atmos. Phys.,* **69,** 81–100.

Pavan, V., and F. J. Doblas-Reyes, 2000: Multi-model seasonal hindcasts over the Euro-Atlantic: Skill scores and dynamic features. *Climate Dyn.,* **16,** 611–625.

Peng, P., A. Kumar, A. G. Barnston, and L. Goddard, 2000: Simulation skills of the SST-forced global climate variability of the NCEP–MRF9 and the Scripps–MPI ECHAM3 models. *J. Climate,* **13,** 3657–3679.

Peng, S., W. A. Robinson, and S. Li, 2003: Mechanisms for the NAO responses to the North Atlantic SST tripole. *J. Climate,* **16,** 1987–2004.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice and night marine air temperature since the late nineteenth century. *J. Geophys. Res.,* **108,** 4407, doi:10.1029/2002JD002670.

Rowell, D. P., 1998: Assessing potential seasonal predictability with an ensemble of multidecadal GCM simulations. *J. Climate,* **11,** 109–120.

Sardeshmukh, P. D., and B. J. Hoskins, 1984: Spatial smoothing on the sphere. *Mon. Wea. Rev.,* **112,** 2524–2529.

——, G. P. Compo, and C. Penland, 2000: Changes of probability associated with El Niño. *J. Climate,* **13,** 4268–4286.

Shukla, J., and Coauthors, 2000: Dynamical seasonal prediction. *Bull. Amer. Meteor. Soc.,* **81,** 2593–2606.

Smith, C. A., and P. D. Sardeshmukh, 2000: The effect of ENSO on the intraseasonal variance of surface temperatures in winter. *Int. J. Climatol.,* **20,** 1543–1557.

Straus, D. M., and J. Shukla, 1997: Variations of midlatitude transient dynamics associated with ENSO. *J. Atmos. Sci.,* **54,** 777–790.

Van den Dool, H. M., and Z. Toth, 1991: Why do forecasts for ''near normal'' often fail? *Wea. Forecasting,* **6,** 76–85.

Wallace, J. M., G.-H. Lim, and M. L. Blackmon, 1988: Relationship between cyclone tracks, anticyclone tracks, and baroclinic waveguides. *J. Atmos. Sci.,* **45,** 439–462.

WASA Group, 1998: Changing waves and storms in the Northeast Atlantic? *Bull. Amer. Meteor. Soc.,* **79,** 741–760.

Whitaker, J. S., and P. D. Sardeshmukh, 1998: A linear theory of extratropical synoptic eddy statistics. *J. Atmos. Sci.,* **55,** 237–258.

Zhang, Y., and I. M. Held, 1999: A linear stochastic model of a GCM's midlatitude storm tracks. *J. Atmos. Sci.,* **56,** 3416–3435.