On the Level and Origin of Seasonal Forecast Skill in Northern Europe

ÅKE JOHANSSON

Swedish Meteorological and Hydrological Institute, Norrkoping, Sweden

ANTHONY BARNSTON

Climate Prediction Center, National Centers for Environmental Prediction, NWS/NOAA, Washington, D.C.

SURANJANA SAHA

Environmental Modeling Center, National Centers for Environmental Prediction, NWS/NOAA, Washington, D.C.

HUUG VAN DEN DOOL

Climate Prediction Center, National Centers for Environmental Prediction, NWS/NOAA, Washington, D.C.

(Manuscript received 22 August 1996, in final form 20 May 1997)

ABSTRACT

This study examines the level and origin of seasonal forecast skill of surface air temperature in northern Europe. The forecasts are based on an empirical methodology, canonical correlation analysis (CCA), which is a method designed to find correlated patterns between predictor and predictand fields. A modified form of CCA is used where a prefiltering step precedes the CCA as proposed by T. P. Barnett and R. Preisendorfer. The predictive potential of four fields is investigated, namely, (a) surface air temperature (i.e., the predictand field itself), (b) local sea surface temperature (SST) in the northern European area on a dense grid, (c) Northern Hemisphere 700-hPa geopotential height, and (d) quasi-global SST on a coarse grid. The design is such that four contiguous predictor periods (of 3 months each) are followed by a lead time and then a single predictand period (3 months long). The shortest lead time is 1 month and the longest is 15 months. The skill of the CCA-based forecasts is estimated for the 39-yr time period 1955–93, using cross-validated hindcasting. Skill estimates are expressed as the temporal correlation between the forecasts and the respective verifying observations.

The forecasts are most skillful in the winter seasons with a secondary weaker skill maximum during summer. During winter the geopotential height field produces the highest skill scores of the four predictor fields. The dominant predictor pattern of the geopotential height field is confined to the predictor period that is closest to a preceding core winter season and resembles the North Atlantic Oscillation (NAO) teleconnection pattern. The time series of the expansion coefficients of this dominant predictor pattern correlates well with a low-pass filtered time series of an NAO index. The obtained skill is similar to what is found in the United States, both with regard to seasonal distribution and level of skill. The origin of skill is however different. In the United States it is the El Niño–Southern Oscillation (ENSO) with its predominantly interannual character that is the main source of skill in winter. In northern Europe it is instead the NAO that contributes the most, and especially the lower frequency part of the NAO (periods between 4 and 10 yr).

Spatially sparse station data of surface pressure extending back to the middle of the nineteenth century suggests a nonstationarity in the NAO behavior. The implications of this nonstationarity for the obtained results of this study is briefly discussed. Because finely resolved field data are not readily available for this earlier period, the level of skill realizable for that period using a pattern relationship technique such as CCA remains an open question.

1. Introduction

Long-range forecasting, especially seasonal forecasting, has received renewed interest in recent years. Much of this interest is associated with the realization that the dominant phenomenon on the seasonal-to-interannual timescale, El Niño–Southern Oscillation (ENSO), has useful predictive skill at forecast lead times of up to 1 yr (see, e.g., Barnston et al. 1994; Latif et al. 1994). Long-range forecasting is thus closely related to predicting the evolution of low-frequency phenomena such as ENSO. Such phenomena are to a large extent a result of interactions between the atmosphere, the oceans, and the land surfaces. Efforts aimed at modeling these interactions and the concomitant low-frequency phenom-

Corresponding author address: Dr. Åke Johansson, Swedish Meteorological and Hydrological Institute, 601 76 Norrköping, Sweden. E-mail: ake.johansson@smhi.se

ena have accelerated in recent years, and considerable work has been devoted to the construction and utilization of comprehensive coupled atmosphere–ocean–land general circulation models (GCMs). However, there is still no complete prediction system with such a GCM and a fully coupled data assimilation procedure, even though work along these lines is now under way at several institutions around the world.

An alternative to using coupled dynamical models is to use empirical methods. Empirical long-range forecasting is an old activity (Rossby 1941; Namias 1953) with documented, but rather modest, skill (Nicholls 1980; Nap et al. 1981; Livezey 1990). Through access to the relatively newly available global datasets, and through the development of new techniques to extract and understand the structures of the low-frequency part of the variability, this activity has advanced considerably in recent years (see, e.g., Van den Dool 1994). The literature on empirical methods for seasonal forecasting is voluminous. While a comprehensive review of this subject exists for the Tropics (Hastenrath 1991), no such review is yet available for the extratropics.

Even though the techniques used in empirical forecasting are statistical, useful information about the nature and dynamics of the low-frequency phenomena present can be obtained. As pointed out by Lorenz (1956), the success of empirical methods is dependent on the fact that the low-frequency phenomena are governed by physical laws that presumably do not change with time. These laws are telling us that the past, present, and future behavior of the atmosphere are related. Thus, the dynamics itself suggest that predictions based on empirical methods should have some skill, and furthermore, the empirical skill levels should get higher and higher the more predictable the atmosphere is by dynamical methods.

The purpose of this paper is to explore the level and origin of skill of seasonal forecasts for northern Europe. We use an empirical method for this purpose. The skill obtained by empirical methods can be regarded as a basic reference, or benchmark, to which the skill levels of coupled dynamical model forecasts should be compared. One reason for using statistical models is that they, in general, require orders of magnitude less computations than do coupled dynamical models. It is therefore possible to extensively explore the method's characteristics on the available historic data. Furthermore, empirical methods have the advantage of providing the possibility of finding relationships between different fields that can give valuable insight into the governing dynamics as well as contribute to improved conceptual models of low-frequency phenomena. The use of a hierarchy of both empirical and dynamical models of varying degrees of complexity is believed to be mutually beneficial for the progress of our understanding of the coupled climate system.

The rationale for use of data and empirical method is given in section 2. The data are described in section 3, while the methodology of the empirical technique is presented in section 4. After presenting the outline of the experiments, a discussion of the average skill in northern Europe is found in section 5. The characteristic geographical as well as temporal skill distributions are presented in sections 6 and 7. An interpretation of the origin of the skill is given in section 8, which is then in section 9 contrasted with the findings of studies of the magnitude and origin of skill in the United States by the same methods. The results are discussed and conclusions are drawn in section 10.

2. Rationale

We know from predictability studies that the atmosphere has an intrinsic range of predictability that is strongly scale dependent. Phenomena characterized by large spatial scales have a longer time range of predictability than those of smaller spatial scales (Lorenz 1969). Similarily, phenomena characterized by large temporal scales (low-frequency phenomena) are more predictable than those of shorter temporal scales (Van den Dool and Saha 1990). Furthermore, a phenomenon's spatial and temporal scales are somewhat correlated. Even though these fundamental results from predictability studies are primarily concerned with synoptic scales and NWP models, they should nevertheless be instrumental in guiding the design of empirical seasonal forecast models. Consequently, here we are using a statistical technique that is designed to find relationships between patterns of large spatial as well as temporal scales.

There exist several methods for finding correlated patterns between two fields. Bretherton et al. (1992) compare several of these methods and find that a modified version of canonical correlation analysis (CCA) is one of the most accurate techniques available. It is known that CCA is susceptible to sampling variability and therefore subject to large random errors when applied to large data fields. Therefore, in accordance with the findings of Bretherton et al. (1992), we use a modified form of CCA, first introduced by Barnett and Preisendorfer (1987). This modified form is characterized by a prefiltering step in which the data are projected onto a truncated series of empirical orthogonal functions (EOFs). This step considerably decreases the number of degrees of freedom in both predictor and predictand fields and makes the canonical correlations more stable with respect to sampling variability. With these preparations the CCA is provided with patterns of large spatial scales, which presumably are associated with large temporal scales.

In the process of finding such large-scale patterns, we aim to avail ourselves of predictor fields that cover as large a part of the globe as possible, and whose temporal scale is correspondingly long. However, in the search of such predictor fields we do not want to compromise the quality of the data. Clearly, the choice of which

predictor parameters to use has to be determined from well-founded hypotheses on the mechanisms that are thought to be responsible for the variability on the seasonal timescale.

Given the above requirements, and the availability of data, we use three different predictor fields in this study, namely SST, the Northern Hemisphere 700-hPa geopotential height field, as well as the predictand itself. The choice of SST as a predictor is based on the rationale that the ocean, with its slow timescale, constitutes a principal source of predictive information for the atmosphere on the seasonal-to-interannual timescale. The requirements of large spatial and temporal coverage as well as good data availability and quality lead us to use an SST field that is quasi-global in extent (roughly 40°S–60°N) and covers the time period from 1955. However, the work of Van den Dool (1984) and Van den Dool and Nap (1985) demonstrate that even local nearby SST information can contain considerable prognostic value on longer timescales, especially for stations close to the coast of the large continents where the prevailing winds are from the sea. As a large part of the northern European area considered has such coastal locations with climatological onshore winds, we will make a departure from the requirement of large spatial scales for the SST predictand. Thus, in addition to the above described coarse resolution quasi-global SST field, we also use a much finer resolution SST field. It covers the seas adjacent to northern Europe: the northeastern part of the North Atlantic Ocean, the Norwegian Sea, the North Sea, and the Baltic Sea.

The predictand field itself—that is, the surface air temperature at 59 stations in northern Europe—is also used as a predictor. Like the SST field it describes the thermal state at the atmosphere's lower boundary, but unlike the quasi-global SST field it is of regional rather than global or hemispheric extent. This compromise on the spatial extent is defended on the ground that the predictor is the predictand at an earlier time, and thereby most suited for implicitly providing local information.

In addition to predictor fields that describe conditions at the atmosphere's lower boundary, we also wish to have a predictor that contains information on the internal low-frequency variability of the atmosphere itself. We use the 700-hPa geopotential height field for this purpose. It would be preferable to use global fields but, unfortunately, such fields are only available since 1978. We are therefore constrained to use data of a lesser spatial extent. We use gridded data that are available since 1947 and which cover most of the Northern Hemisphere (north of ~20°N).

3. Data

The raw data used in this study consist of monthly mean values and cover the time period 1955–93, that is, 39 yr.

a. Northern European surface air temperature (T)

Monthly mean surface air temperature data from 59 stations in 18 countries in northern Europe are used in this study. Their geographical location is shown in Fig. 1a and some general information about them is listed in Table 1. The stations are located within a circle-shaped area centered around central Sweden. The size of this area is about the same as that of the conterminous United States, that is, around 8×10^6 km^2. The coverage is aimed at being equal-area with a somewhat denser distribution in the Nordic countries. Data for Swedish stations are obtained from the Swedish Meteorological and Hydrological Institute (SMHI), for Finnish stations from the Finnish Meteorological Institute (FMI), and for Dutch stations from the Royal Netherlands Meteorological Institute (KNMI). The data from the remaining stations are taken from the Global Telecommunication System (GTS) as well as from the Carbon Dioxide Information Analysis Center (CDIAC). In general, very few missing data are encountered, as seen in Table 1. This is especially true for the carefully quality controlled data from the three national meteorological institutes. The missing data are replaced through spatial and temporal interpolation.

b. Sea surface temperature (SST)

Monthly mean SST data are taken from the Comprehensive Ocean Atmosphere Data Set (COADS) (Slutz et al. 1985) for the period 1955–79, from ship data from 1980 to 1981 and from blended ship–satellite data from 1982 to 1993 (Reynolds 1988). The data are available in 2° lat by 2° long boxes. To investigate the role of local versus remote SST data we have composed two parts: 1) a fine-resolution SST dataset covering northern European seas and 2) a coarse-resolution SST dataset of quasi-global extent.

1) NORTHERN EUROPEAN SST DATA [SST(L)]

The data cover an area in and adjacent to northern Europe as displayed in Fig. 1b. Mean values of SST in 2° lat by 2° long boxes are given at 82 grid points. We will also refer to this as the local SST data, therefore the acronym SST(ℓ), where the index ℓ stands for local.

2) QUASI-GLOBAL SST DATA [SST(QG)]

The data cover the area between 40°S and 60°N. The original data from the 2° lat by 2° long grid are averaged to form a 10° lat by 10° long grid. This averaged dataset then consists of 268 gridpoint values. However, due to the fact that for certain grid points, especially in the Southern Hemisphere, the basic observations that are used to calculate the box average value are either too sparse in time and space and/or have poor quality, only 235 grid points are actually used. These grid points are

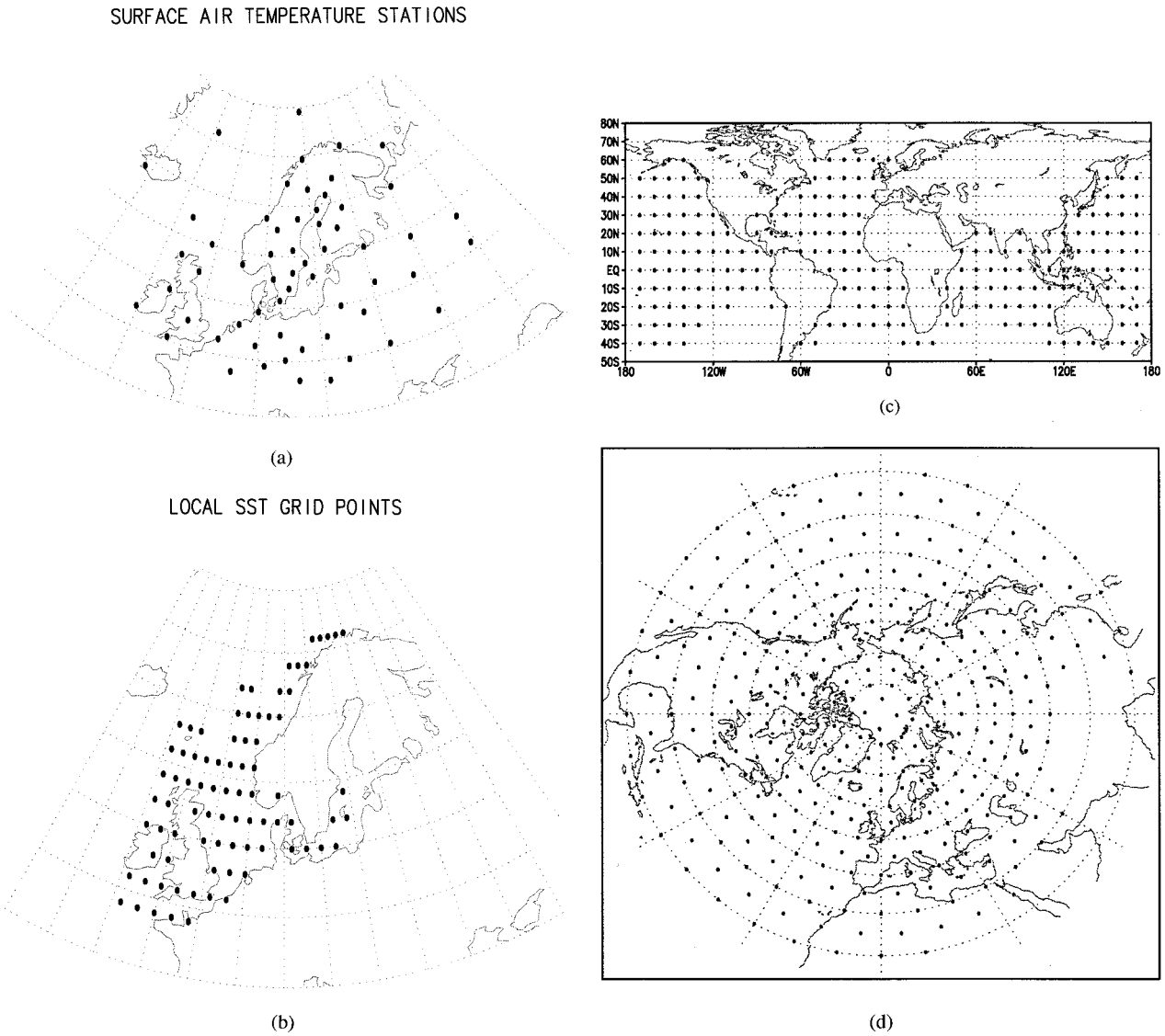

FIG. 1. Geographical distribution of the different predictor and predictand fields. (a) Surface air temperature (T) at 59 stations in northern Europe. (b) Sea surface temperature (SST) values at 82 grid points in and adjacent to northern Europe, SST(l), as represented by the average SST value in 2° lat by 2° long boxes. (c) Quasi-global SST values, SST(qg), at 235 grid points as represented by the average SST value in 10° lat by 10° long boxes. (d) 700-hPa geopotential height (h700) values at 358 grid points.

shown in Fig. 1c. This SST dataset is denoted SST(qg), where qg stands for quasi-global.

c. 700-hPa geopotential height (h700)

This dataset is computed from operational twice-daily analyses of the former National Meteorological Center [NMC, now NCEP (National Centers for Environmental Prediction)]. These analyses are on an approximately equal area 358-point grid covering the Northern Hemisphere from approximately 20°N and northward as shown in Fig. 1d. From these daily analyses monthly mean values are calculated.

4. Methodology

a. Basic definitions

We define a season as a time interval spanning over three contiguous calendar months. We divide the year into 12 overlapping seasons. Thus, there is considerable autocorrelation between neighboring seasons, as they are two-thirds identical. Let t_S be an index for the 12 seasons in a year. This index can have one of the values 1 to 12, where 1 denotes the season DJF (December, January, February), 2 denotes JFM (January, February, March), and so on. The season number thus refers to the number of the middle month of the season as illustrated along the time axis in Fig. 2a.

TABLE 1. Some general information about the surface air temperature stations in northern Europe used in this study.

Name	Country	Latitude	Longitude	Height over MSL (m)	Number of missing data
Jan Mayen	Norway	70.93	−8.67	9	5
Tromsø/Langnes	Norway	69.68	18.92	10	1
Bjørnøya	Norway	74.52	19.02	16	6
Vardø	Norway	70.37	31.10	13	0
Bodø	Norway	67.28	14.42	12	0
Ørlandet	Norway	63.70	9.62	7	0
Oslo/Gardermoen	Norway	60.20	11.08	204	5
Stavanger	Norway	58.88	5.63	8	0
Jokkmokk	Sweden	66.60	19.83	260	0
Haparanda	Sweden	65.83	24.15	5	0
Bjuröklubb	Sweden	64.48	21.58	36	0
Junsele	Sweden	63.68	16.90	210	0
Fjällnäs	Sweden	62.63	12.20	780	0
Falun	Sweden	60.60	15.64	160	0
Stockholm	Sweden	59.33	18.05	44	0
Linköping	Sweden	58.40	15.52	93	0
Fårö	Sweden	57.97	19.35	8	0
Säve	Sweden	57.75	11.87	20	0
Växjö	Sweden	56.87	14.80	166	0
Lund	Sweden	55.70	13.19	50	0
Sodonkylä	Finland	67.37	26.65	180	0
Kajaani	Finland	64.28	27.68	136	0
Vaasa	Finland	63.05	21.77	4	0
Luonetjärvi	Finland	62.40	25.67	140	1
Turku	Finland	60.52	22.27	49	0
Lerwick	Scotland	60.13	−1.18	83	0
Stornoway	Scotland	58.22	−6.32	9	0
Aberdeen	Scotland	57.20	−2.20	72	0
Birmingham	England	52.45	−1.73	99	0
Plymouth	England	50.35	−4.12	27	0
Aldergrove	N Ireland	54.65	−6.22	81	0
Valentia	Ireland	51.93	−10.25	9	0
Reykjavik	Iceland	64.13	−21.93	61	1
Thorshavn	Faroe Islands	62.02	−6.77	39	2
Eelde	Netherlands	53.13	6.58	5	0
Vlissingen	Netherlands	51.45	3.60	10	0
Nancy	France	48.68	6.22	217	2
Schleswig	Germany	54.53	9.55	44	0
Berlin/Tempelhof	Germany	52.47	13.40	50	0
Kassel	Germany	51.32	9.48	233	0
Nürnberg	Germany	49.50	11.08	312	3
Wien/Hohe Warte	Austria	48.25	16.37	209	0
Miscolc	Hungary	48.13	20.80	119	0
Praha/Ruzun	Czech Rep	50.10	14.25	369	0
Warzawa/Okezie	Poland	52.17	20.97	106	0
Wroclaw	Poland	51.10	16.88	120	0
Kaunas	Lithuania	54.87	23.88	75	0
Minsk	Belarus	53.87	27.53	234	0
Kiev	Ukraine	50.40	30.45	179	0
Lvov	Ukraine	49.82	23.95	325	10
Kanin Nos	Russia	68.65	43.30	48	1
Archangelsk	Russia	64.58	40.50	13	5
St Petersburg	Russia	59.97	30.30	2	0
Velikie Luki	Russia	56.38	30.60	98	8
Vologda	Russia	59.28	39.87	116	2
Kirov	Russia	58.65	49.62	164	8
Kazan	Russia	55.78	49.18	64	0
Moscow/Observatory	Russia	55.75	37.57	156	0
Voronezh	Russia	51.70	39.17	164	10

Let $T_O(i, t_S, t_Y)$ denote the observed seasonal mean surface air temperature at station i for season t_S in the year t_Y. The station index i can range from 1 to 59. As the dataset spans the 39-yr period from 1955 to 1993 we let t_Y go from 1 to 39. Let furthermore $T_F(i, t_S, t_Y, t_F)$ denote the forecasted value made at a forecast lead time of t_F, which is verified against the observed value $T_O(i, t_S, t_Y)$. The index t_F denotes the number of months

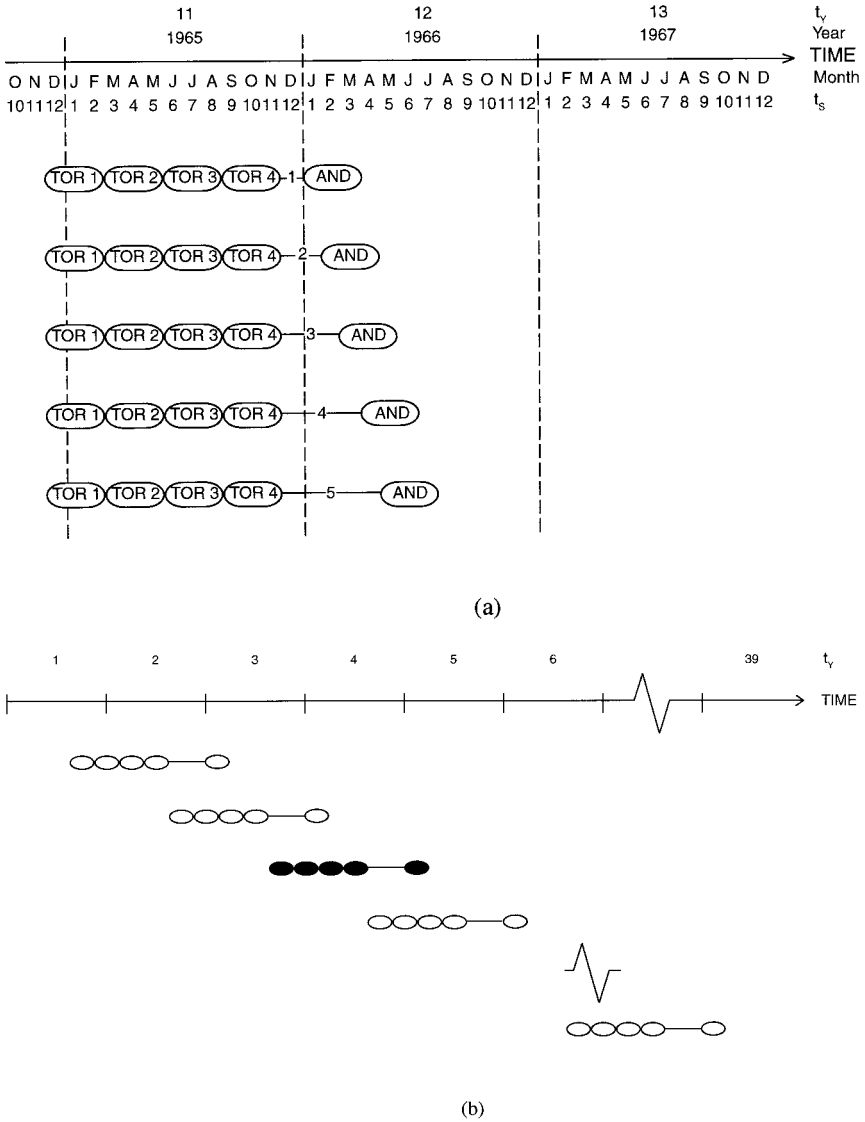

FIG. 2. Graphical illustration of the temporal configuration of the predictor and predictand periods. Each of the predictor and predictand periods is 3-months long (a season). Four nonoverlapping contiguous predictor periods (TOR1–TOR4) are followed by a lead time and then a single predictand period (AND). This whole sequence is called a torand-sequence. (a) Configuration of torand-sequences for forecasts with progressively longer lead times. The lead time increment is 1 month. The shortest lead time is 1 month and the longest is 15 months. (b) Configuration of all the torand-sequences used for forecasting the JFM season ($t_S = 2$) in year 5 at a lead time of 4 months. This targeted torand-sequence is marked in black and is not used in the model building process.

between the end of predictor data to the beginning of the predictand data (targeted season) as shown, by example, in Fig. 2a.

b. Predictor and predictand periods and fields

In Fig. 2a a graphical illustration is shown of the temporal occurrence and extent of the predictor and predictand periods for progressively longer forecast lead times. Each of the predictor and predictand periods is a 3 month (seasonal) average. Four nonoverlapping contiguous predictor periods (TOR1–TOR4) are followed by a lead time, t_F, which is ranging from 1 month up to 15 months, which in turn is followed by a single predictand period (AND). This whole sequence will here be denoted a *torand-sequence*, TAS. A torand-sequence is uniquely determined by the season and year of the predictand period and the forecast lead time, that is, TAS(t_S, t_Y, t_F).

The empirical model building process consists of

TABLE 2. The combination of predictor fields for the various experiments performed to predict surface air temperature. The predictor fields are denoted T for surface air temperature, SST(ℓ) for local sea surface temperature, h700 for the 700-hPa geopotential height, and SST(qg) for quasi-global sea surface temperature. The SST(qg) field is furthermore divided into various ocean domains, which are described in Table 3. The code for the experiment designator is as follows. The numbers indicate the predictor fields used: 1 for T, 2 for SST(ℓ), 3 for h700, and 4 for SST(qg). In the experiments that include SST(qg) as a predictor there is also a symbol in the decimal place, which represents which ocean domain is used.

Experiment designator	Field weight of the four predictor fields				Designator for ocean domain of SST (qg)
	T	SST(ℓ)	h700	SST(qg)	
1	1	0	0	0	—
2	0	1	0	0	—
3	0	0	1	0	—
4.G	0	0	0	1	G
4.1	0	0	0	1	1
4.2	0	0	0	1	2
4.3	0	0	0	1	3
4.4	0	0	0	1	4
4.5	0	0	0	1	5
4.6	0	0	0	1	6
12	1	1	0	0	—
13	1	0	1	0	—
23	0	1	1	0	—
123	1	1	1	0	—

finding coupled patterns between the data in the four predictor periods and the data in the predictand period (the target season) for each and every one of all the torand-sequences. This is done in the framework of performing *hindcasting*, that is, in the construction of the empirical model for a specific torand-sequence, TAS(t_S^o, t_Y^o, t_F^o), all other TAS(t_S^o, t_Y, t_F^o) are used, not only those from antecedent times ($t_Y < t_Y^o$) but also those of later origin than the targeted one ($t_Y > t_Y^o$) are made available. However, because of the overwhelming importance of the yearly cycle, we use for each year only that torand-sequence that has the same target season (t_S^o) and forecast lead time (t_F^o) as the targeted one. An illustration of this procedure is shown in Fig. 2b for the case when the JFM season (season 2) in year 5 is forecasted at a lead time of 4 months. A total of 36 torand-sequences are here used for the construction of the empirical model for this specific torand-sequence.

c. Data preprocessing

The raw data as such is not very suitable to be used directly in a canonical correlation analysis. We therefore process the data in a series of steps, whereby we seek to refine robust large-scale features that hopefully are also characterized by slow timescales. The procedure follows closely that used by Barnston (1994) (hereafter referred to as B94) and is graphically summarized in the upper part of Fig. 2 of B94.

Each predictor and predictand data element is first standardized so that all elements are weighted equally

TABLE 3. Description of the various ocean domains that form subsets of the SST(qg) dataset.

Designator for ocean domain of SST (qg)	Geographical name of ocean domain	Latitude interval	Number of grid points
G	Quasi-global oceans	40°S–60°N	235
1	Extratropical North Atlantic Ocean	30°N–60°N	24
2	Extratropical North Pacific Ocean	30°N–60°N	33
3	Atlantic Ocean	40°S–60°N	65
4	Pacific Ocean	40°S–60°N	129
5	Tropical oceans	30°S–30°N	139
6	Tropical Pacific Ocean	30°S–30°N	86

in producing a correlation-based CCA. The standardization ensures that all locations have equal opportunity to participate in the predictive patterns, regardless of their latitude- and longitude-dependent interannual variances. The standardization is performed in the framework of hindcasting as described above in section 4b. A data element, e.g., T_O, is thus standardized (T'_O) according to

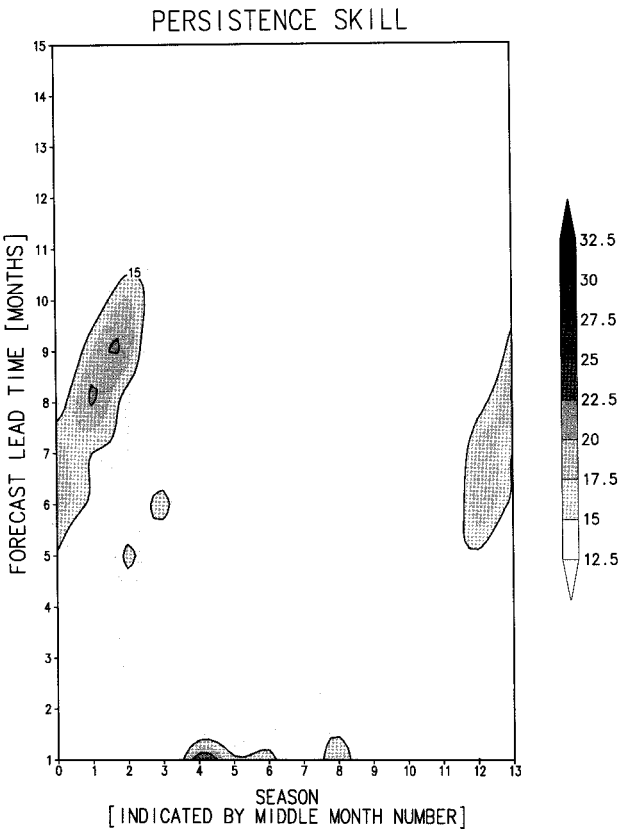

FIG. 3. The area average cross-validated anomaly correlation coefficient according to (5) for persistence forecasts of surface air temperature. The target season is on the abscissa and the forecast lead time is on the ordinate.

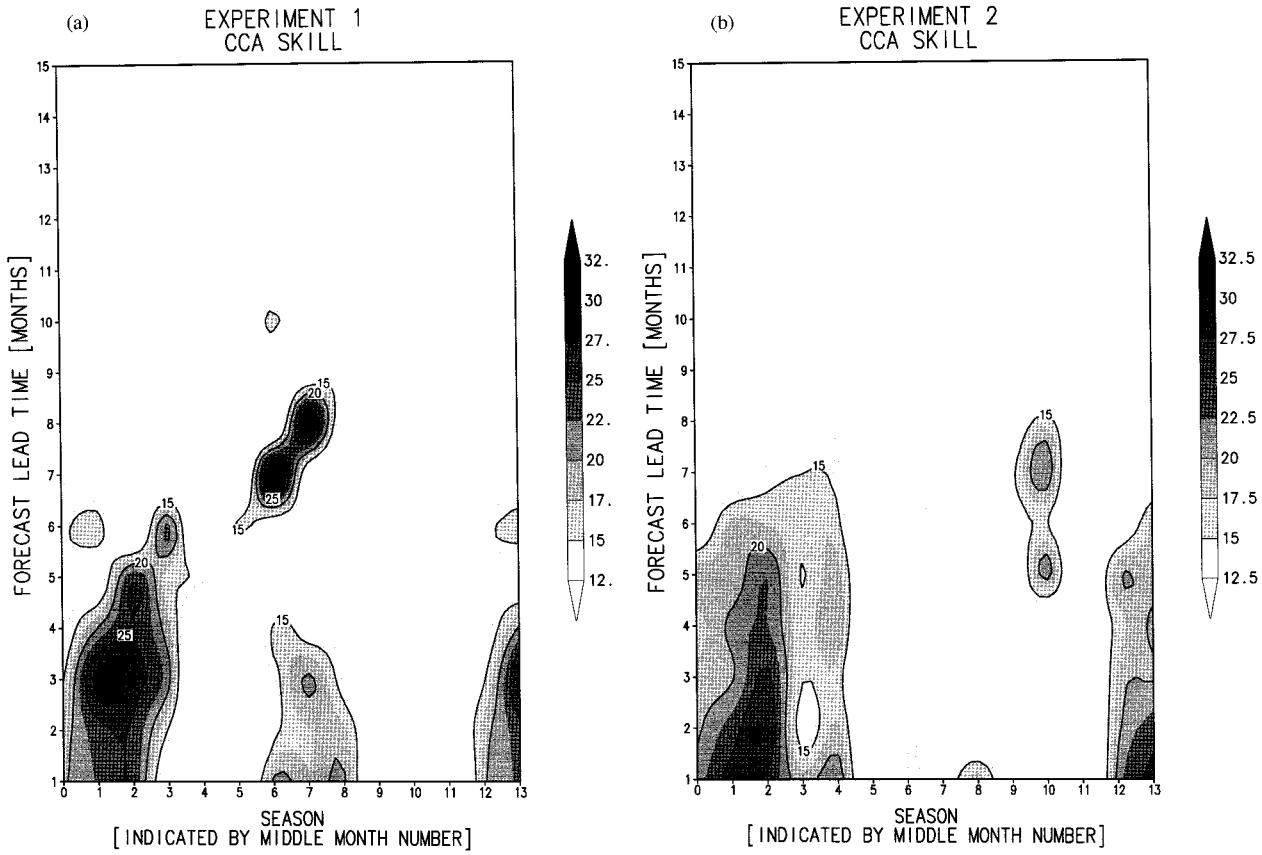

FIG. 4. As in Fig. 3 except for CCA forecasts with (a) the predictand itself as the only predictor (experiment 1), (b) local SST as the only predictor (experiment 2), (c) the geopotential height field at 700 hPa as the only predictor (experiment 3), and (d) quasi-global SST as the only predictor (experiment 4.G).

$$T'_O(i, t_S, t_Y) = \frac{T_O(i, t_S, t_Y) - \overline{T_O(i, t_S)}}{\sigma(i, t_S)}, \qquad (1)$$

where

$$\overline{T_O(i, t_S)} = \frac{1}{(N - n)} \sum_{\substack{t_Y = n \\ t_Y \neq t_Y^o}}^{N} T_O(i, t_S, t_Y) \qquad (2)$$

is the mean value and

$$\sigma(i, t_S) = \sqrt{\frac{1}{(N - n)} \sum_{\substack{t_Y = n \\ t_Y \neq t_Y^o}}^{N} [T_O(i, t_S, t_Y) - \overline{T_O(i, t_S)}]^2} \qquad (3)$$

is the standard deviation. The year t_Y^o is withheld in (2) and (3) because it is the target prediction year. Due to the length of data required for each forecast (see Fig. 2b), which depends upon forecast lead time, some of the years at the beginning and sometimes at the end of the 39-yr time period cannot be used. This fact is taken into account as the summation starts in year $n > 1$ and ends in year $N \leq 39$.

Second, we establish an interfield weighting, which adjusts the relative weighting of two or more predictor fields, as wholes, relative to one another. If this step were not carried out, the relative weights would be proportional to the relative number of elements (grid points or stations) in the respective fields.

The predictor and predictand datasets are thereafter filtered through an orthogonalization procedure as suggested by Barnett and Preisendorfer (1987). Separate EOF analyses are thereby performed and only the leading modes are retained. Within the predictor pre-orthogonalization, the predictor fields from four consecutive prior periods (forming a temporal sequence) enter into the process together, but as individual predictors, producing an extended (in time as well as across predictor fields) EOF analysis. Here the relative weighting of the predictor fields, as wholes, is of crucial importance, for it determines to what extent the resulting orthogonal components contain the influence of one field versus another. The EOF mode truncation point is based partly on Monte Carlo experiments that determine the statistical separability of the eigenvalue curve from that resulting from random input data (Preisendorfer 1988). Additionally, truncation occurs before any mode whose eigenvalue is less than one percent of that of the leading mode. These criteria normally imply that five to six modes are retained.

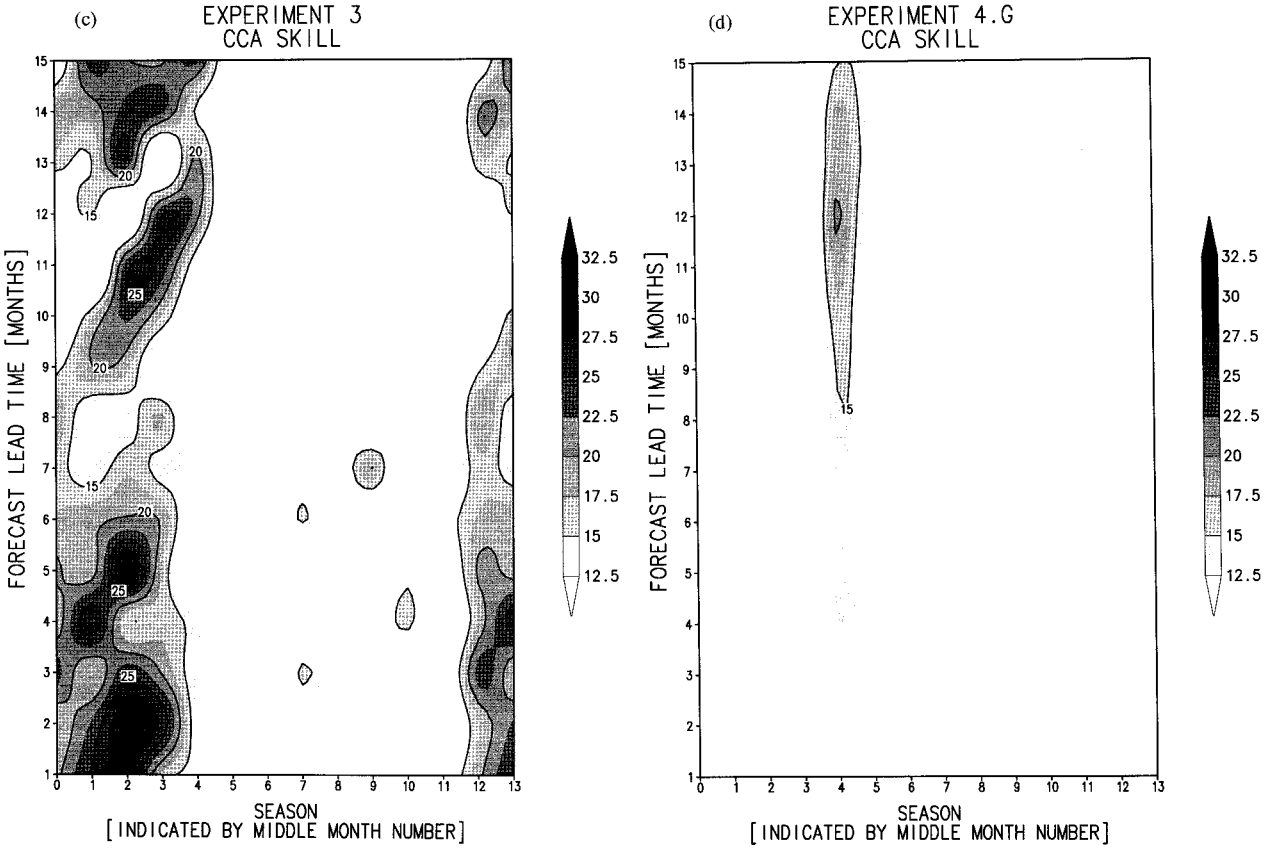

FIG. 4. *(Continued)*

d. Canonical correlation analysis

The prefiltered data are then subjected to a regular, classical CCA as described by, for example, Glahn (1968). We apply the same truncation rules for the canonical modes as described above in section 4c in the prefiltering procedure. The retained canonical modes, usually five to six, are then used for prediction.

e. Assessment of forecast skill

An exhaustive version of cross validation is used in order to avoid artificial skill and thus obtain as representative skill estimates as possible. The hindcast procedure described above in section 4b is thereby repeated for all available years in the dataset and a mean skill of all these forecasts is calculated. This exhaustive version of cross validation is described and discussed in Michaelsen (1987) and has since been used in a large number of statistical forecast studies (see, e.g., B94 for references).

The measure of forecast skill used is the temporal correlation between observed and forecasted temperature anomalies. We thus define a cross-validated anomaly correlation coefficient, ACC, as

$$\mathrm{ACC}(i, t_S, t_F) = \frac{\sum_{t_Y=n}^{39} T'_F \cdot T'_O}{\sqrt{\sum_{t_Y=n}^{39} T'^2_F \cdot \sum_{t_Y=n}^{39} T'^2_O}}. \quad (4)$$

The starting year in the summation, n, is either 2, 3, or 4 as forecasts for the earliest years are not possible as there are no data for the required preceeding predictor periods (see Fig. 2b).

5. Average skill over northern Europe

In the following we shall describe a set of experiments that assess the quality of seasonal forecasts in northern Europe. The experiments differ by the choice of predictor variables and their spatial resolution and distribution. A summary of the various experiments is given in Table 2. Note that for the experiments with quasi-global SST as a predictor we have made a division into several subareas, which are described in Table 3.

To get an overall estimate of the skill level over the entire northern European area, we consider the average skill of all the 59 stations according to:

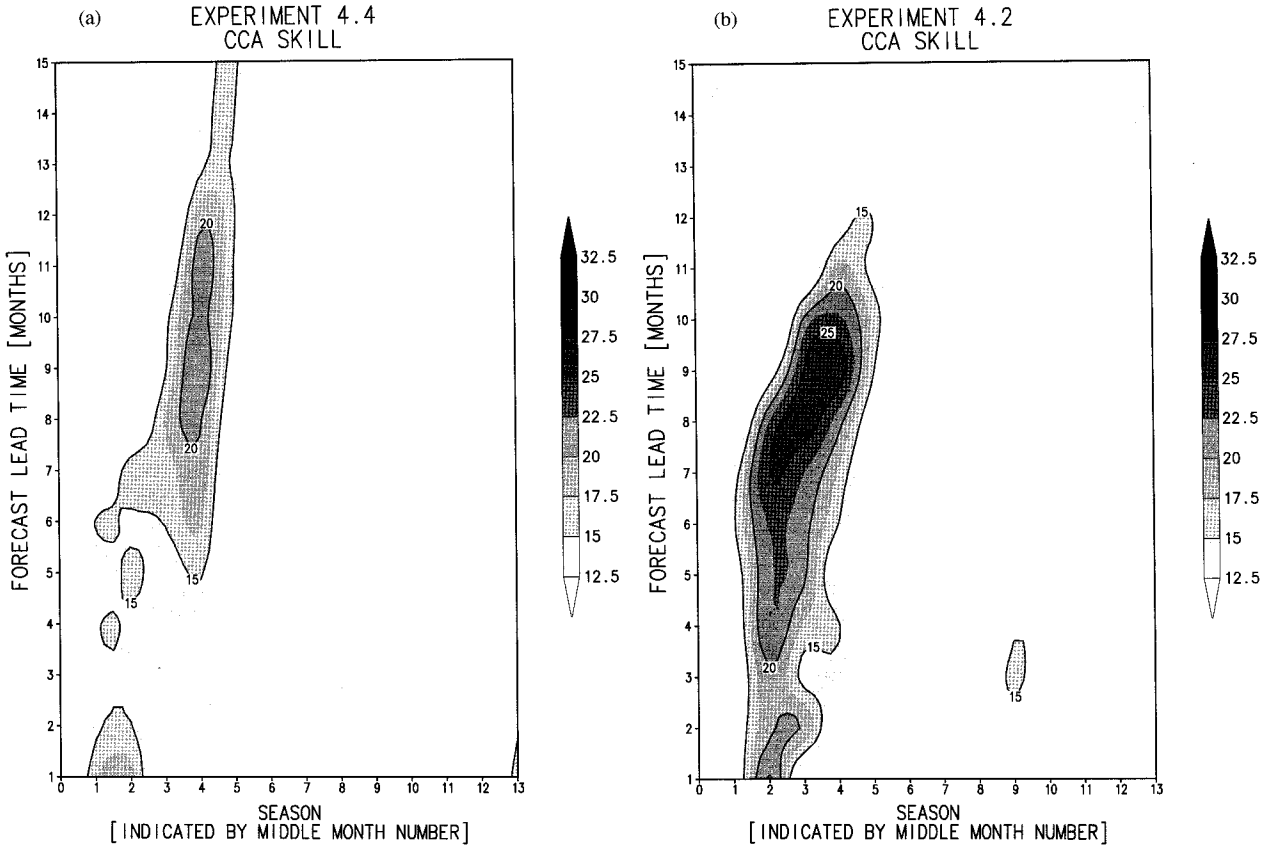

FIG. 5. As in Fig. 4d except that the quasi-global SST is restricted to only (a) Pacific Ocean as described in Table 3 (experiment 4.4), (b) extratropical North Pacific Ocean as described in Table 3 (experiment 4.2), and (c) tropical Pacific Ocean as described in Table 3 (experiment 4.6).

$$\overline{\mathrm{ACC}}(t_S, t_F) = \frac{1}{59}\sum_{i=1}^{59} \mathrm{ACC}(i, t_S, t_F). \qquad (5)$$

The measure is emphasizing the contribution from stations in Scandinavia where we have the highest density of stations. In the discussion below, the area average cross-validated anomaly correlation coefficient according to (5) will be displayed as a two-dimensional contour plot with the target season on the abscissa and the forecast lead time on the ordinate.

a. Persistence

Persistence is used as the basic reference, or benchmark, to which the skill levels of the CCA forecasts are compared. For persistence forecasts we have a modified torand-sequence from that displayed in Fig. 2. There is only one predictor period in this case; that is, TOR1–TOR3 is removed and TOR4 becomes the only predictor period, TOR. The cross-validated version of hindcasting used for estimating persistence forecast skill implies that for each specific torand-sequence we use a damped persistence determined by the autocorrelation coefficient between predictor and predictand periods as determined from all the other torand-sequences.

As seen in Fig. 3, the highest persistence skills are found for forecasts of the winter season for lead times of 6–10 months. According to our definitions of predictor, predictand, and forecast lead periods (see Fig. 2), a lead time of 9 months implies a temporal offset of 12 months between the center of the predictor period and the center of the predictand period. Earlier and later seasons apparently do not matter for persistence of anomalies toward next winter. In particular the 1-month lead (autumn to winter) persistence is near zero. Correlations in excess of 0.20 are found for the core winter seasons DJF and JFM. The slantwise shape of the skill maxima implies that it is the previous JFM season (t_S = 2) that is of pivotal importance as it alone apparently contains predictive information for almost all of the seasons in the next winter (t_S = 12, 1, 2, 3). In contrast to the long-lead nature of the skill maxima in winter, the summer (t_S = 4–8) has noticeable persistence skill only at the shortest leads.

It is interesting to contrast the above result with the skill of classical month-to-month persistence forecasts (e.g., Craddock and Ward 1962; Van den Dool and Nap 1981; Van den Dool and Livezey 1984; Vedin et al. 1991). Despite being compiled on a monthly basis and

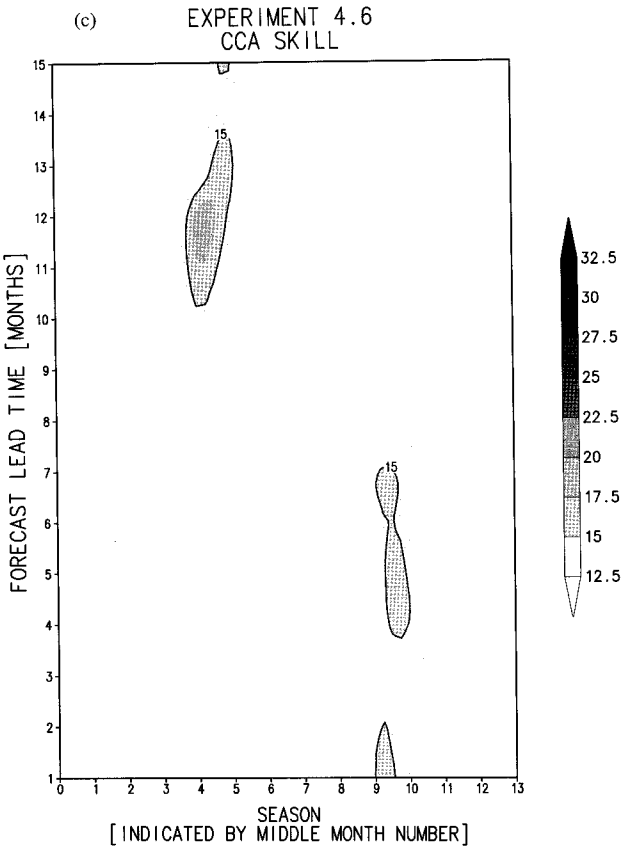

FIG. 5. *(Continued)*

with zero lead time ($t_F = 0$), they are similar to the above seasonal long lead results in that they possess two maxima during the year, one in the winter and one in the summer seasons.

b. Experiments with only one predictor field

1) SURFACE AIR TEMPERATURE (THE PREDICTAND ITSELF) AS THE ONLY PREDICTOR

As seen in Fig. 4a, there are two local skill maxima during the year when the predictand itself at earlier times is used as the predictor in CCA. The largest skill values occur during winter, while a secondary maximum occurs during summer. Two distinct minima during the intervening transition seasons, spring and autumn, are clearly distinguishable. There is virtually no skill beyond a lead time of 9 months. We believe that a physically sound skill distribution should show a monotonic decrease of skill with increasing forecast lead time, since there is more and more information available with decreasing lead time. The ''return of skill'' that can be seen in the summer is therefore a questionable feature. One reason that this occurs in this experiment may be the choice of design of predictor periods. As seen from Fig. 2 the chosen design implies that old, potentially important information is not made available for short lead predictions. If this assumption is correct it would imply that it is possible to obtain increased skill in summer with a more suitable design of predictor periods. An alternative explanation for the ''return of skill'' is that the high ACCs for long leads in summer are due to sampling errors and therefore a result of chance.

It is appropriate to contrast Fig. 3 with Fig. 4a as both these prediction experiments are based on surface air temperature as the only predictor. Both experiments exhibit a similar seasonal dependence with maximum skill in winter and summer and minima in between. They differ sharply, however, in that the CCA forecasts are much more skillful at short leads. We attribute this to the fact that the CCA technique uses both a longer predictor sequence and is capable of finding more refined correlated patterns between predictors and predictand. Specifically, we note that the important JFM season from the previous winter, which is so pivotal for local persistence forecasts, is included in the predictor sequences for short leads. The fact that persistence beats CCA at long leads in winter might be due to the fact that surface air temperature in the more distant predictor periods (beyond the last winter) do not contain any predictive information and therefore only introduces noise to the CCA technique, so as to confuse any signal coming from the most recent JFM season. It may also be that the prefiltering step in the present CCA procedure removes too much of the local effects evident in Fig. 3.

2) LOCAL SST AS THE ONLY PREDICTOR

Figure 4b displays the skill from the experiment where the predictor is SST from the waters in and adjacent to northwest Europe (local SST) as shown in Fig. 1b. The skill distribution and magnitude are reminiscent of the case with surface air temperature as the only predictor. This is reasonable as both of these predictors represent the thermal conditions at the lower boundary of the atmosphere in the region considered. Note especially that the forecast skill does not extend beyond $t_F = 8$ months.

3) GEOPOTENTIAL HEIGHT FIELD AT 700 HPA AS THE ONLY PREDICTOR

In this experiment the only predictor is the geopotential height field at 700 hPa covering most of the Northern Hemisphere north of 20°N as shown in Fig. 1d. The skill maximum is now confined to the winter seasons, as seen in Fig. 4c, and with magnitudes of similar order as when surface air temperature is the only predictor. The skill tends, however, to be high for longer leads and is not appreciably diminished at a lead of 15 months. The geopotential height field thus seems to contain information capable of producing meaningful winter forecasts more than 1 year in advance. As discussed above in section 5b(1), we believe that the vacillating

skill distribution with forecast lead time is a result primarily of the specific design of predictor periods.

4) Quasi-global SST as the only predictor

In the following experiments the only predictor is quasi-global SST, SST(qg), with a spatial resolution as shown in Fig. 1c. The skill scores when the full SST(qg) field is used are shown in Fig. 4d. Noticeable forecast skill values are confined to only one target season, namely MAM. In contrast with the experiment with surface air temperature as predictor but in a similar fashion to the one with the geopotential height field as predictor, we find detectable skill at long lead times. Maximum skill occurs for $t_F = 12$, which corresponds to correlations with SST that existed between one and two years earlier. Quasi-global SST thus appears to have some, albeit very slight, potential as a predictor for longer lead times in the spring season. Even though it is difficult to make a clear comparison (e.g., different verifying areas and time periods), this result seems to agree with some dynamical modeling results (e.g., Branković et al. 1994), which find a tendency for the skill to be highest in the spring season. The reason why the maximum occurs in spring is not rigorously clarified but there have been some suggestions (see, e.g., Branković et al. 1994). The main conclusion from our experiments, however, is that quasi-global SST contains relatively little predictive information for northern Europe. This is the case both in comparison with the three other predictor fields, as well as in comparison with the large and dominant role it has over the United States (B94).

Having established the small influence of quasi-global SST as a predictor, it might seem of little value to investigate how the different ocean basins contribute to the observed skill levels. The paramount importance of SST, and in particular tropical Pacific SST, for seasonal forecasting in the United States (e.g., B94) and the suggestions of an ENSO signal in Europe (Fraedrich and Müller 1992; Fraedrich 1994), makes such an investigation nevertheless of interest. For this purpose we divide the world oceans into six subbasins as described in Table 3. In Fig. 5 we present the resulting skill levels from three of these six experiments, of which all three involve the Pacific Ocean. The other three experiments produce very low skill levels and therefore are not shown. The skill scores when the whole Pacific Ocean is used as predictor are shown in Fig. 5a. The skill distribution is somewhat similar to the experiment with the full SST(qg) field as shown in Fig. 4d in that the skill is confined to the spring seasons, $t_S = 4$–5, for long leads. However, we also find noticeable additional skill at short leads in winter. Dividing the Pacific Ocean into subbasins, an extratropical Northern Hemisphere part (Fig. 5b) and a tropical part (Fig. 5c), reveals that the short lead winter skill comes primarily from extratropical North Pacific SST while the long lead spring skill comes from tropical SST.

When contrasting Figs. 4d, 5a, 5b, and 5c, it is evident that skill scores are not additive in a simple way. For example, the use of Pacific Ocean SST alone produces higher skill scores as compared to when SST from all of the three major oceans is used. This fact indicates that the Atlantic Ocean, despite being much closer to the target area, does not seem to contain much predictive information, and furthermore, seems to introduce noise into the CCA procedure.

It is of interest to contrast this result with other studies concerning the role of Atlantic SST for long-range forecasting in the northern European area. Palmer and Sun (1985) have studied relationships between SST anomalies in the extratropical north Atlantic, in particular the area southeast of Newfoundland, and the Northern Hemisphere circulation. Their results suggest that these relationships are not so useful for seasonal prediction, as (i) the lagged correlations are largest for the atmosphere leading the ocean and (ii) the correlations are greatest for small lag (around a month). Lead–lag relationships between Atlantic SST modes and atmospheric modes have recently been reevaluated by Deser and Timlin (1995), who find that the SST lags the atmospheric circulation by only 2–3 weeks. Thus, these findings support our result that North Atlantic SST is not so useful for predictions of the atmosphere on the seasonal timescale at lead times greater than 1 month.

The reason that the use of North Pacific SST produces relatively skillful forecasts may be related to the longer term, decadal atmosphere–ocean variability that has been observed in and around the North Pacific Ocean (e.g., Trenberth and Hurrell 1994). Furthermore, the North Pacific area is affected by an extratropical component of ENSO, which also might contribute to the obtained skill levels. The skill scores obtained using only North Pacific SST are of similar magnitude as found in experiment 2 (where only local SST is used), but with the important difference that the large skill scores are concentrated at longer lead times (around 9 months). Surprisingly, ENSO itself does not appear to have a strong direct influence, when judged from the low skill scores shown in Fig. 5c.

5) Discussion of the experiments with only one predictor field

There is a tendency in most of the experiments with only one predictor for the best skill scores to be found during the winter seasons. The lowest skill scores are generally found during the transition seasons, spring and autumn, except for the case with Pacific Ocean SST as predictor. There is a weaker secondary maximum in the summer.

An interesting difference in the behavior of forecast skill scores as a function of forecast lead time is clearly distinguishable among the experiments. While the experiments with surface air temperature and local SST as predictors have usable skills that do not extend much

beyond a lead time of 9 months, the predictors with more remote, hemispheric to global coverage (the geopotential height field and Pacific Ocean SST) have recognizable skill, with much longer leads, in some cases out to at least 15 months.

The best skill scores overall are found for the geopotential height field as a predictor. Here the high skill values are confined to the winter seasons where the skill values decay only slightly with forecast lead time.

c. *Experiments with several predictor fields*

The logical next step is to investigate the impact on the distribution and magnitude of skill of having several predictor fields working together. The overall impression is that adding another predictor field produces similar, or at best only modestly increased, skill scores compared to using only the best of the involved predictors alone. A reason for this may be that the predictors contain somewhat redundant information. The surface air temperature at the Norwegian coast and on the islands in the Norwegian Sea is, for example, highly correlated with the SST in the Norwegian Sea. Another reason is that correlations are, of course, not additive. The correlation between two predictor fields together and the predictand is in a nontrivial way related to the correlations between each of the predictors individually and the predictand.

The three experiments where the geopotential height field is included as one of the predictors have very similar skill distributions and furthermore they are similar to the experiment with the height field as the only predictor. Therefore it seems natural to conclude that the height field is the dominant source of predictive skill and that the other fields only contribute marginally to increase the skill. The experiment that produces the highest overall skill scores is the one with surface air temperature and geopotential height field as predictors, experiment 13, for which the corresponding skill scores are shown in Fig. 6.

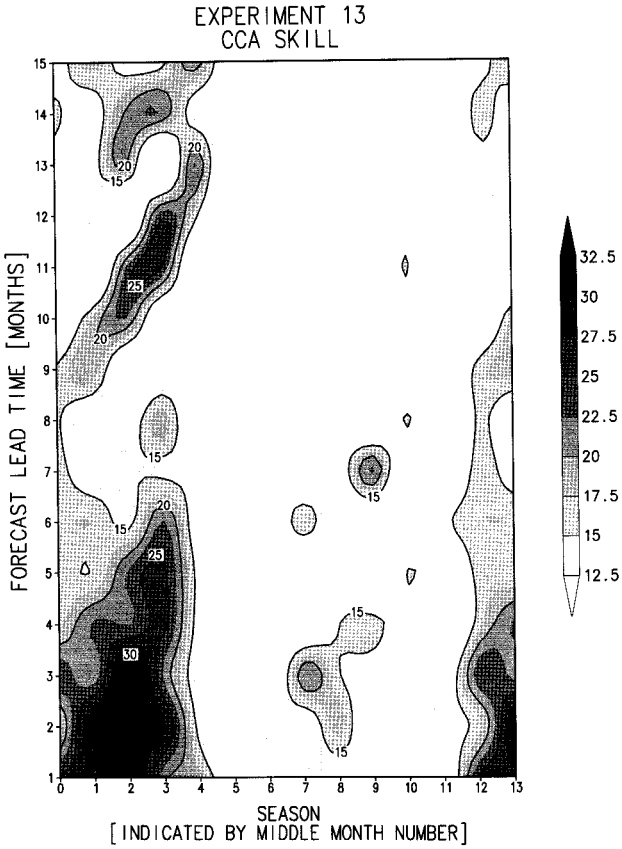

FIG. 6. As in Fig. 4a except with a predictor combination of surface air temperature and the geopotential height field at 700 hPa (experiment 13).

6. Geographical distribution of skill

So far we have considered only the mean skill for all of the 59 stations. The geographical distribution of the skill is, of course, also of interest. From the area average skill diagram of section 5 we can conclude that the maximum skill scores most often occur for short leads in the winter seasons. In Fig. 7 we therefore display the geographical skill distribution for the JFM season for the shortest forecast lead time, $t_F = 1$ month. Shown are the results from three of the most skillful experiments, namely, those with surface air temperature and 700-hPa geopotential height field as sole predictors, experiments 1 and 3, respectively, as well as the experiment with the best combination of predictors: surface air temperature and the height field, experiment 13. The case with surface air temperature as only predictor (Fig. 7a) has higher skill values in a band from the northern North Sea to southern Russia (correlation skill values above 0.30 are shaded progressively more darkly). The two cases with the height field included as predictor (Figs. 7b,c) have skill values in excess of 0.30 in a larger area that covers the whole southeastern half of the domain. Maximum values are above 0.45 and are found in a belt stretching from southeastern Sweden over the Baltic states and Poland to farther eastward toward Moscow. The quality of the forecasts decreases in all cases toward the northwest with insignificant skill over Iceland.

7. Temporal distribution of skill

To obtain a perhaps more intuitive appraisal of the quality of the seasonal forecasts, we present Fig. 8. It shows a year-to-year comparison between an observed and forecasted time series for the same target season (JFM) and forecast lead time ($t_F = 1$ month) as in the previous section and for experiment 13 with the best combination of predictors: surface air temperature and the height field. The station considered is Fårö on the Swedish island of Gotland, as this station has an anom-

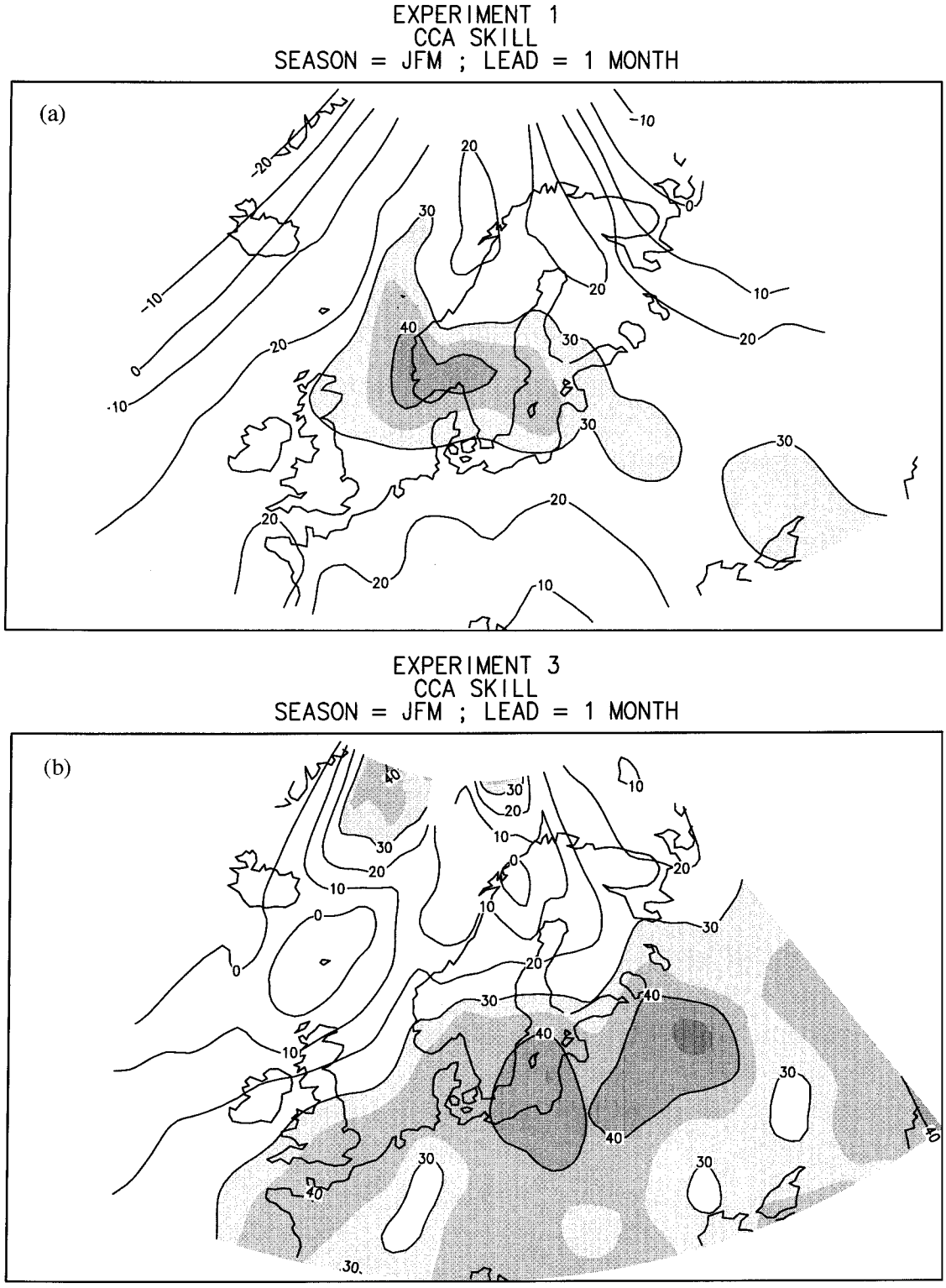

FIG. 7. The geographical skill distribution for the CCA forecasts for the JFM season and for the shortest forecast lead time, t_F = 1 month. The CCA forecasts are from the experiments with (a) surface air temperature and (b) 700-hPa geopotential height field as only predictors, respectively, as well as (c) for the experiment with the best combination of predictors: surface air temperature and the geopotential height field.

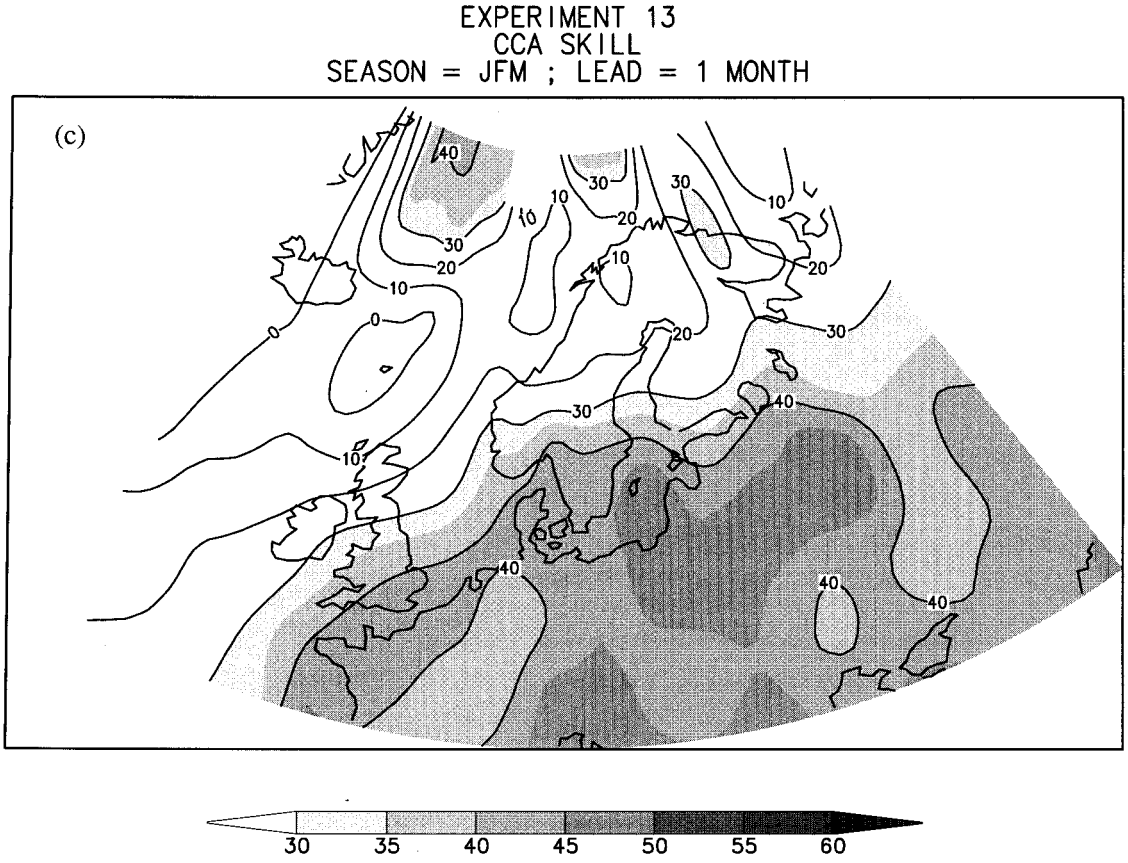

FIG. 7. (*Continued*)

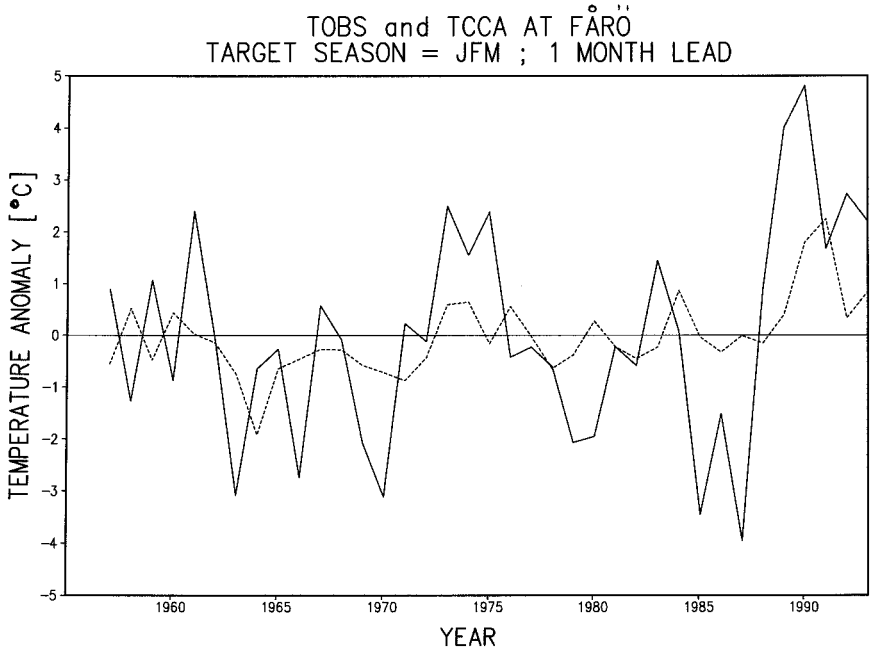

FIG. 8. Time series of observed (solid line) and forecasted (dashed line) anomalies at Fårö, Sweden, for the JFM season and for the shortest forecast lead time, t_F = 1 month. The CCA forecasts are for the experiment with the best combination of predictors: surface air temperature and the geopotential height field (experiment 13).

aly correlation coefficient of 0.48, and thus proved to be one of the most skillful stations.

We first note that the CCA forecasts underestimate the amplitude of the anomalies. This is a common feature of statistical forecasting techniques that seek to minimize the root-mean-square error. There are cases with large forecast errors, but in general the curves tend to follow each other. Examples of time periods where the CCA technique have some skill in capturing the observed evolution is the cold period of 1963–70 as well as the recent sequence of warm winters.

8. On the origin of skill

The CCA technique provides some tools that may help shed some light on the question of the origin of the skill. These tools consist of the canonical eigenvalues and eigenvectors that are calculated in the CCA procedure as well as the time series of the expansion coefficients of the canonical eigenvectors. Here we will call them ''the diagnostics.'' The canonical eigenvectors of the predictand and predictors are perhaps the most important diagnostics. The geographical distribution and magnitude of these eigenvectors give important information on which phenomena might be responsible for the relationships that the CCA technique is able to find and subsequently use to form its forecasts.

Examination of the diagnostics from the different experiments reveals a rather coherent picture, which enables us to limit the discussion to a single experiment without missing any essential relationships. Based on the results from the experiments described in section 5 we will consider therefore only the case with the geopotential height field together with surface air temperature as predictors (experiment 13), as it was found to be the most skillful case in an overall sense. Furthermore, we will focus on a target season in winter (JFM, $t_S = 2$) and a forecast at a short lead ($t_F = 1$ month). It is very helpful in the following discussion to be fully aware of the temporal configuration of the predictor and predictand periods, as displayed symbolically in Fig. 2. Even if the forecast lead time is only 1 month, the predictor fields extend more than one year back in time and furthermore consist of four separate consecutive seasons. Of special importance is the fact that the first predictor period (TOR1) is the winter season DJF one year earlier than the target season, JFM.

In Fig. 9 the canonical eigenvector of the geopotential height predictor field is shown for each of the four predictor periods. Here we consider only the dominant, first mode. It is clear that the eigenvector for the first predictor period (TOR1) has substantially larger amplitude than those for the three other periods. Of all the predictor periods the first one is farthest removed in time from the target period, but it is a winter season (DJF of the previous winter). We will have reasons to emphasize the significance of this in the following. The spatial pattern is characterized by one center in the Iceland–south Greenland area and a band of opposite sign along 40°N with maxima in southern Europe and the southeastern United States. This pattern is reminiscent of the North Atlantic Oscillation (NAO) as described, for example, in chart 2 in the work by Walker and Bliss (1932), in the idealized NAO in Fig. 1 of Wallace and Gutzler (1981) and in Figs. 2a and 2.1 of Barnston and Livezey (1987). The physical significance of this pattern is that in the phase characterized by an anomalously deep Icelandic low, the relatively warm and humid air that normally blows from the Atlantic Ocean in over Europe is strengthened and gives rise to milder than normal winters. The reverse situation occurs in the opposite phase. We also see hints of this pattern in predictor periods 2 and 4—that is, the spring and autumn seasons.

Now consider Fig. 10, which shows the canonical eigenvector of the surface air temperature field as a predictor for the first of the four predictor periods (TOR1). We do not show the other three predictor periods as they are all weaker in a similar fashion to the height field. The spatial pattern of the eigenvector suggests that this pattern is associated with the height field distribution, that is, with an NAO type of pattern. Here we make the interpretation that the height field is the fundamental governing predictor and that the surface air temperature, as well as the local SST field (not shown), mainly contain the height field predictor information, but in an indirect diluted form.

The corresponding canonical eigenvector for the predictand is shown in Fig. 11. The pattern is characterized by a rather flat field with high values over most of the continental areas and a sharp decrease toward the northwest. The fact that Figs. 10 and 11 are similar but not identical would superficially indicate that the CCA forecasts are similar to persistence forecasts. However, we have shown in the above that persistence forecasts (Fig. 3) have much less skill than the CCA forecasts (e.g., Fig. 6). Figure 11 is a close approximation to the forecast for the next winter, but Fig. 10 may not be close at all to what was observed in the previous winter. It is the strength of CCA to distill out of the observations available at the previous winter, those components that should be persisted to the next winter.

The interpretation that the NAO is the main source of predictive skill as well as the main cause of the observed anomalies warrants a consideration of the NAO in some detail. A way to describe the temporal evolution of the NAO is to study the time series of a single number that represents the phase and amplitude of a predetermined NAO pattern. Here we form such a single number by calculating a very simple NAO index (NAOI) corresponding to the anomalous geostrophic wind (in m s^{-1}) as calculated from the seasonal mean surface pressure difference between Stykkishólmur, Iceland, and Ponta Delgada, the Azores. As these two stations are located almost on the same longitude, the calculated geostrophic wind is close to having only an east–west

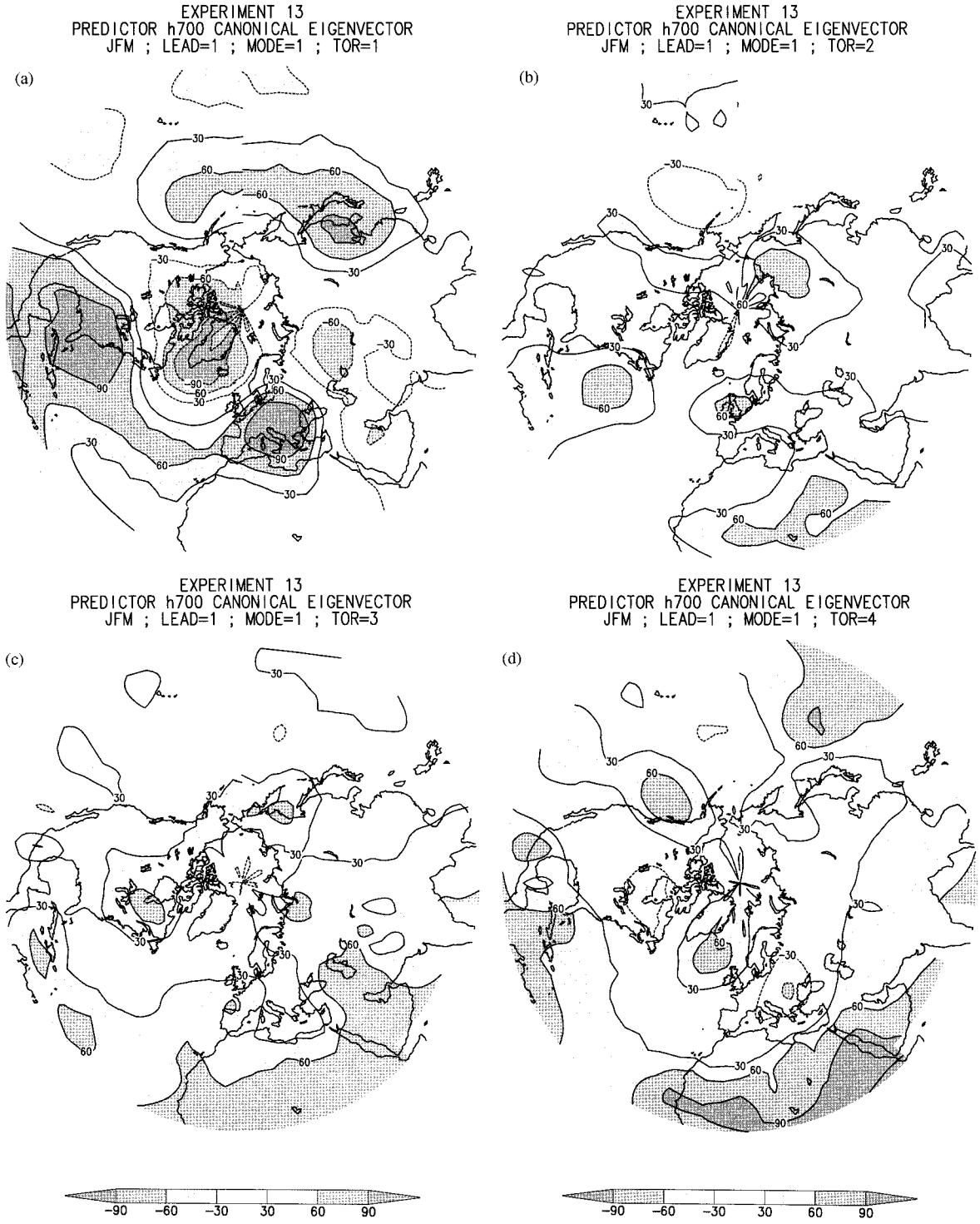

FIG. 9. The canonical eigenvectors of the geopotential height field as predictor for each of the four predictor periods (a)–(d). The eigenvectors are for experiment 13 and for the target season JFM and for forecast lead time t_F = 1 month.

component. A positive value of this index indicates above normal westerly winds blowing from the Atlantic Ocean onto Europe. The reason we use this specific form of an NAOI is that it represents a physically easily interpretable quantity. One way of displaying the NAOI that emphasizes its seasonality is shown in Fig. 12. Two graphs are shown: Fig. 12b presents a smoothed version of the raw data displayed in Fig. 12a. The filtering is

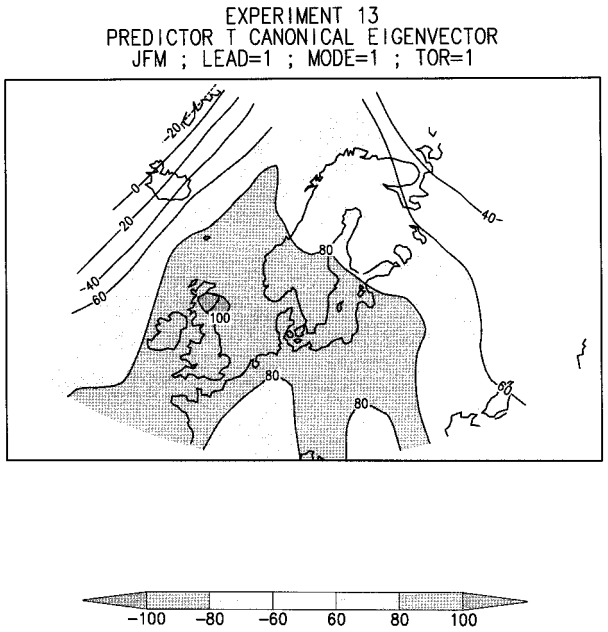

FIG. 10. As in Fig. 9 except for surface air temperature as the predictor and only for the first predictor period.

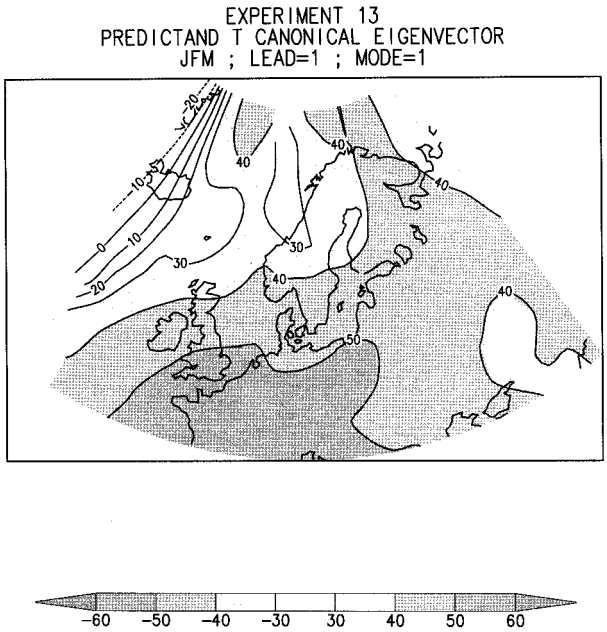

FIG. 11. As in Fig. 10 except for the canonical eigenvector of the predictand.

performed in the vertical direction, that is, for each season a low-pass filter with seven weights (1, 3, 5, 6, 5, 3, 1) is applied to neighboring years in order to remove fluctuations with periods less than 4 yr (Hurrell 1995). Two features are of interest. First, the NAOI has large values almost exclusively in the winter season. Second, the NAOI has a tendency to retain the same sign several winters in a row, sometimes for a whole decade and longer. An interesting aspect is the repetition of large values of the same sign winter after winter irrespective of the fact that the index almost disappears in the intervening summer and can even have opposite sign during spring and autumn.

Further support for the notion that the NAO in the winter seasons is the dominant phenomenon responsible for the observed surface air temperature anomalies as well as the obtained predictive capability is provided by the set of graphs shown in Fig. 13. All graphs refer to the JFM season. In Fig. 13a we compare the time series of the raw NAOI (solid line) with the simultaneously observed surface air temperature anomalies at the station Fårö (dashed line). The reason for choosing this station is the same as explained in section 7. There is a very high degree of similarity between these two quantities (correlation coefficient of 79%) suggesting that temperature anomalies in the region are strongly related to the NAO. This result is in line with findings in earlier work (e.g., Walker and Bliss 1932; van Loon and Rogers 1978). In Fig. 13b a comparison between the time series of the smoothed NAOI (solid line) and the expansion coefficient of the first mode of the canonical predictor eigenvector for the JFM target season for 1 month forecast lead time (dashed line) are shown. The similarity between the curves (correlation coefficient of 67%) suggests that the CCA technique finds and exploits parts of the low-frequency variations associated with the NAO, corresponding roughly to variations on the 4–10-yr timescale. A similar comparison but using the raw, unsmoothed, NAOI (not shown) shows much less similarity (correlation coefficient of 46%). Finally, in Fig. 13c we compare the time series of the expansion coefficient of the first mode of the canonical predictor eigenvector for the JFM target season for 1-month forecast lead time (solid line) and the forecasted surface air temperature anomalies at the station Fårö (dashed line). The similarity between the curves is very high (correlation coefficient of 86%), which indicates the dominant role of the first canonical mode in shaping the forecast. We have already shown the accompanying comparison between the observed and forecasted surface air temperature for the same station in Fig. 8a.

In the previous paragraph we emphasize the role of the low-frequency part of the NAO for the CCA predictions. A simpler way to exploit this low-frequency NAO variability for predictions is to use lagged correlations between an NAO index and surface air temperature. A cross-validated skill estimate for such a forecast based on the NAO index defined above and the surface air temperature at the station Fårö is shown in Fig. 14a. As a reference we show the skill distribution for persistence forecasts in Fig. 14b. The high skill at fairly long leads in the winter season is distributed slantwise, corresponding to a relationship between temperature at the target winter and the NAO index in the previous winter. Expanding Fig. 14a to even longer fore-

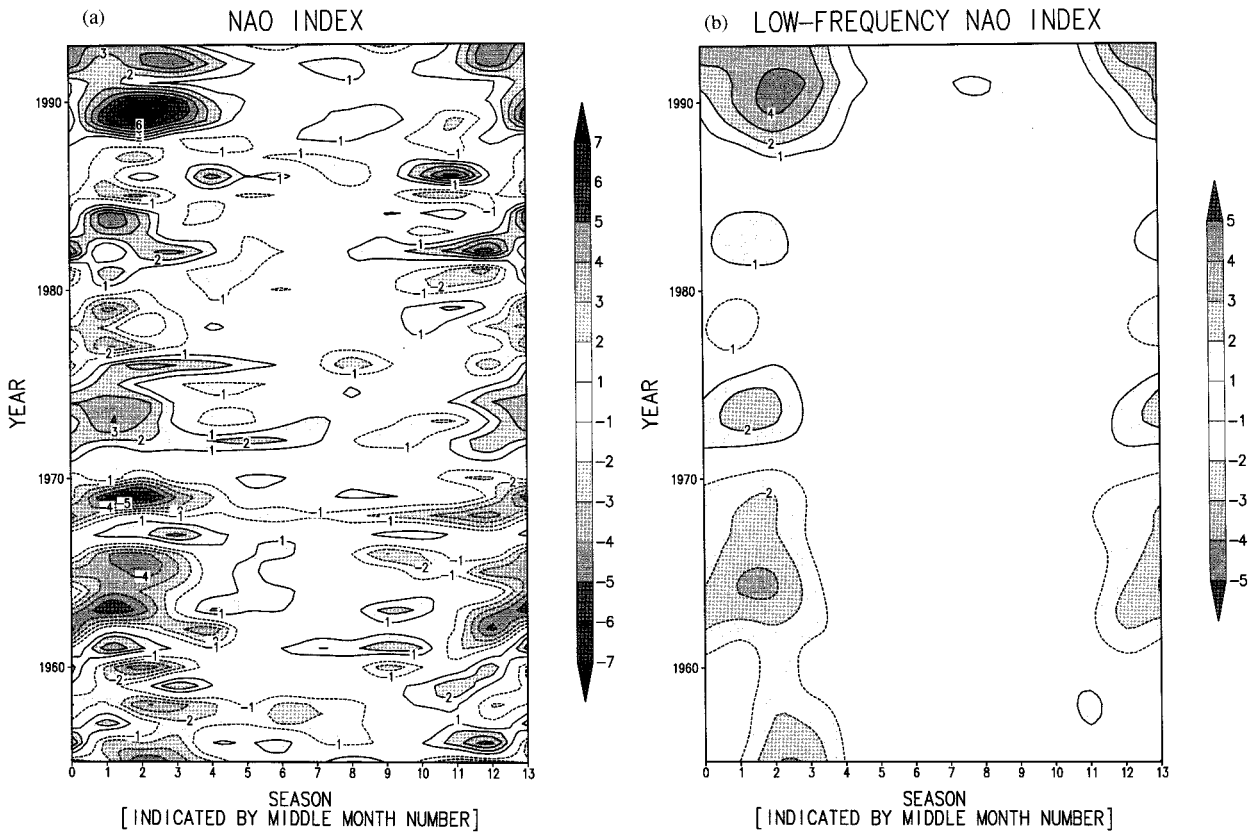

FIG. 12. An index of the North Atlantic Oscillation (NAOI) corresponding to the anomalous geostrophic wind (in m s^{-1}) as calculated from the seasonal mean surface pressure difference between Stykkishólmur, Iceland (65°05′N, 22°44′W), and Ponta Delgada, the Azores (37°44′N, 25°42′W). As these two stations are located almost on the same longitude the calculated geostrophic wind is close to having only an east–west component. The season is on the abscissa and the year is on the ordinate. (a) Raw NAOI data. (b) Filtered NAOI data. The filtering is performed in the vertical direction; that is, for each season a low-pass filter with seven weights (1, 3, 5, 6, 5, 3, 1) is applied to neighboring years in order to remove fluctuations with periods less than 4 yr.

cast lead times (not shown) reveals that there are no further high skill (>30%) forecasts, at, for example, 2 or 3 winter leads. As seen in Fig. 14b pure persistence has a similar skill distribution but is less skillful in the winter season. In Fig. 14c we finally present the corresponding skill obtained from CCA forecasts (experiment 13). Contrasting Figs. 14a and 14c it is evident that the CCA technique produces winter season forecasts that are more consistent with respect to forecast lead time. After all, the skill should monotonically decrease with increasing lead. This is not the case in either of these graphs but the CCA forecasts are closer to having this property. Furthermore, the CCA forecasts have generally a higher skill level that extends all the way to 15 months lead time. We thus conclude that the CCA technique is superior in exploiting the predictive information associated with the low-frequency part of the NAO than simple correlations.

9. Comparison of seasonal forecast skill between northern Europe and the United States

The 59 stations in northern Europe considered in this study cover an area that is about the same size as the contiguous 48 states of the United States. The similarity in areal size and their location in the extratropical Northern Hemisphere makes a comparison of skill levels between the two regions appropriate.

The skill scores of seasonal forecasts of surface air temperature over the United States as obtained from the CCA technique are shown in Fig. 15. These forecasts are produced in a similar way to the ones described above for northern Europe, that is, the same CCA methodology is used with similar predictor fields covering the same 39-yr period. Two of the three predictor fields used are identical (quasi-global SST and 700-hPa geopotential height fields), whereas the third (the surface air temperatures) is of course from stations in the United States. The figure shows the result from the best combination of predictors, which is composed of all of the above three predictors, with the quasi-global SST field weighted twice as much as the others. This is done to emphasize and take advantage of the influence ENSO has on short-term climate anomalies over the United States. Figure 15 should thus be compared to Fig. 6b, which shows the best result for northern Europe.

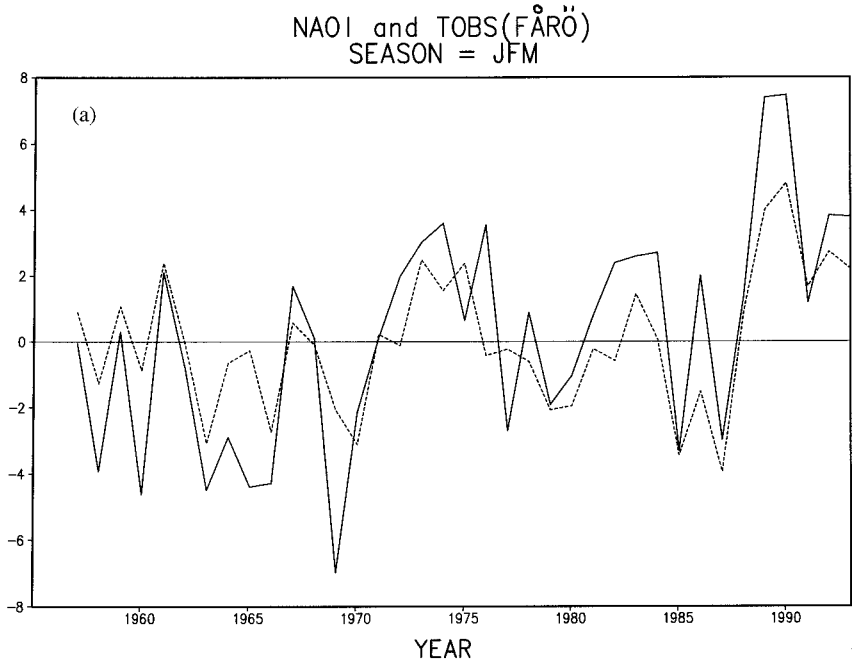

FIG. 13a. Time series of the unfiltered NAO index, NAOI (solid line), and the observed surface air temperature anomalies at the station Fårö (dashed line) both for the JFM season.

There is a striking resemblance between the skill score distributions with a maximum during winter to late winter and a weaker, secondary maximum during summer to late summer. However, the most remarkable finding is that the skill scores are of similar magnitude. In the absence of a strong ENSO signal in Europe, it has otherwise been presumed that the skill scores over Europe should be considerably lower than what is found for the United States.

The sources of the skill are, however, very different

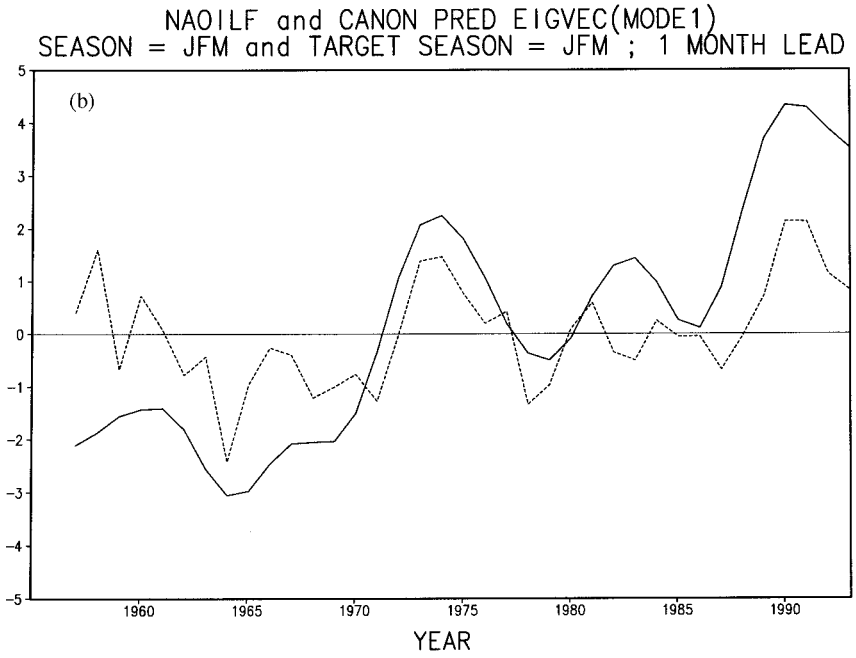

FIG. 13b. Time series of the filtered NAOI for the JFM season (solid line) and the expansion coefficient of the first mode of the canonical predictor eigenvectors for the JFM target season and for 1 month forecast lead time (dashed line) for experiment 13.

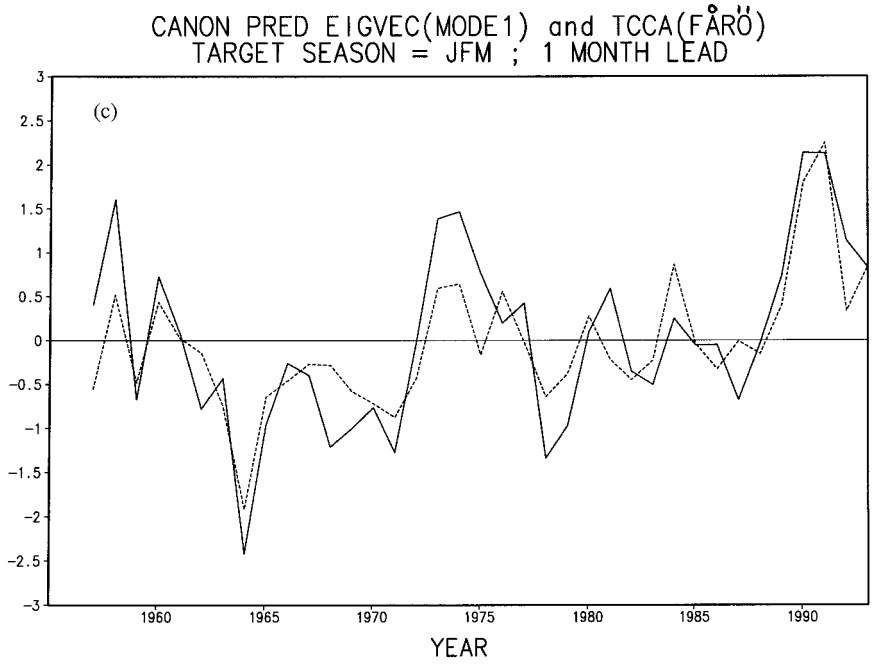

FIG. 13c. Time series of the expansion coefficient of the first mode of the canonical predictor eigenvectors (solid line) and the forecasted surface air temperature anomalies at the station Fårö (dashed line), both for the JFM target season and for 1-month forecast lead time (experiment 13).

for the two regions. For the United States, skill is based primarily on ENSO-related SST from the tropical Pacific Ocean (see B94), whereas in Europe this study indicates that skill is based on the variations of the geopotential height field associated with the North Atlantic Oscillation (NAO). The NAO varies typically at a lower frequency than ENSO (Hurrell and van Loon 1994). Therefore, the predictability in northern Europe has more of an interdecadal flavor, which often manifests itself as a winter-to-winter persistence, in contrast to the predictability in the United States, which is more interannual in character.

10. Discussion and conclusions

This study is an extension of the work presented in B94, where an empirical technique is used to investigate short-term climate predictive skill in the Northern Hemisphere. We use the same basic statistical methodology, the CCA technique, but have restricted ourselves in several other aspects. First of all, we consider only forecasts of seasonal means, as we believe a season to be the shortest averaging period that allows an acceptable signal-to-noise ratio and the longest averaging period that does not mix different times of year too much. Second, we forecast only surface air temperature, and third we confine the interest to the northern European area. However, we have also expanded in that the number and quality of the predictor fields have been increased. Two of the four predictor fields are identical to the ones used in B94, that is, the geopotential height field at 700 hPa from 20°N northward and quasi-global SST fields on a 10° lat by 10° long grid. The third predictor field, which is also the predictand field, is the surface air temperature at 59 stations in northern Europe. The larger number of stations in the area considered and the higher quality of their time series is the basic difference and improvement compared to the European temperature study in B94. The fourth predictor field, which is not used in B94, is a finer resolution SST field on a 2° lat by 2° long grid covering the waters adjacent to northern Europe. The time span considered, the 39-yr period from 1955 to 1993, is a couple of years longer than that considered in B94.

In order to establish which combination of predictors produces the highest overall skill scores, we calculate an area-averaged temporal anomaly correlation coefficient (ACC) for the northern European area. In displaying and intercomparing this quantity as a function of target season and forecast lead time for various combinations of predictors, we obtain a rather clear picture of how the different predictors contribute to the skill. We defer this discussion to the next paragraph and now discuss only the dependence on target season and forecast lead time. We do this for the case of the combination of geopotential height field and surface air temperature as predictor fields, as this combination proves to give the best results. For this case we can discern a pronounced seasonal dependence of the skill. The seasonal variation of skill occurs almost exclusively as a function of the target season, as the dependence on the temporal lead is almost negligible. Specifically, the winter target

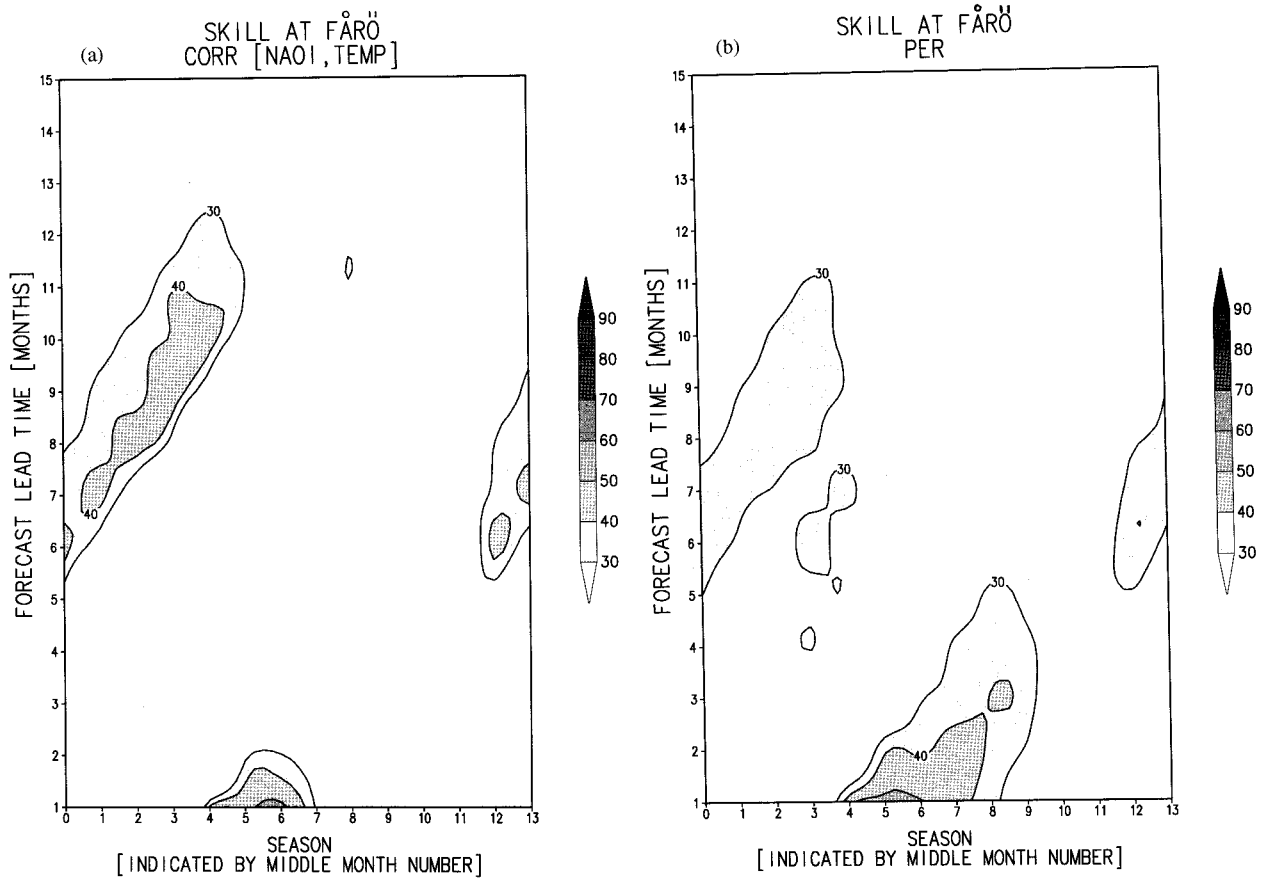

FIG. 14. The cross-validated correlation coefficient for forecasts of surface air temperature at the station Fårö based on (a) lagged correlations with NAOI, (b) persistence forecasts, and (c) CCA forecasts according to experiment 13.

seasons prove to be the most skillfully predicted. A much smaller secondary maximum in skill occurs during the summer seasons, with practically no skill in the intervening transition seasons. The slow decrease of skill with increasing lead time in the winter season, implies the possibility of meaningful seasonal forecasts for winter at long leads. It is interesting to compare the above-described results for northern Europe with those for the best predictor combination for CCA forecasts over the conterminous United States. Such a comparison reveals a surprising similarity, both with respect to overall skill levels as well as to the seasonal cycle in skill with two maxima, one in the winter and one in the summer season. Furthermore, the similarity continues with the winter maximum being the strongest, as well as the slow decrease of skill with increasing lead time. Despite these similarities the origin of skill appears to be quite different for the two areas, as will be discussed below.

From studying the area-averaged ACC over northern Europe for the various combinations of predictors, we find that most of the skill originates from the geopotential height field. Some additional skill, especially in the summer season, comes from the surface air and sea temperature fields in the northern European area. Examination of the predictor and predictand canonical eigenvectors reveals that a pattern reminiscent of the North Atlantic Oscillation (NAO) is a major factor explaining the obtained skill levels. This notion is also supported by the high correlation between an NAO index (NAOI) and the observed surface air temperature at a typical northern European location in winter. Furthermore, we find that the CCA technique apparently captures the lower frequency variability of the NAOI, roughly corresponding to the 4–10-yr timescale, and is less skillful in predicting the interannual variations in the NAOI. Therefore we hypothesize that the NAO is the dominant predictable low-frequency phenomenon (at least in the winter season) on timescales ranging from a season to a decade in the northern European area. It is of course a strong point of CCA to capture a low-frequency filtered version of the NAOI.

The findings of the present study can be contrasted with those of the less detailed European temperature study in B94. The sparsity of Scandinavian stations and the inclusion of southern Europe in B94 cause results to be markedly different from those found here. Skill in B94 for northern Europe is much more modest than

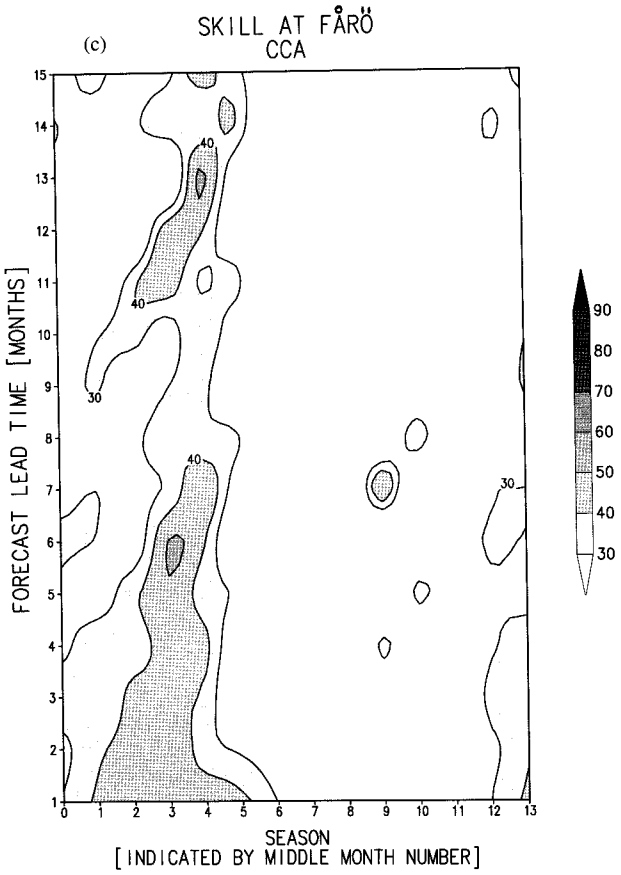

FIG. 14. *(Continued)*

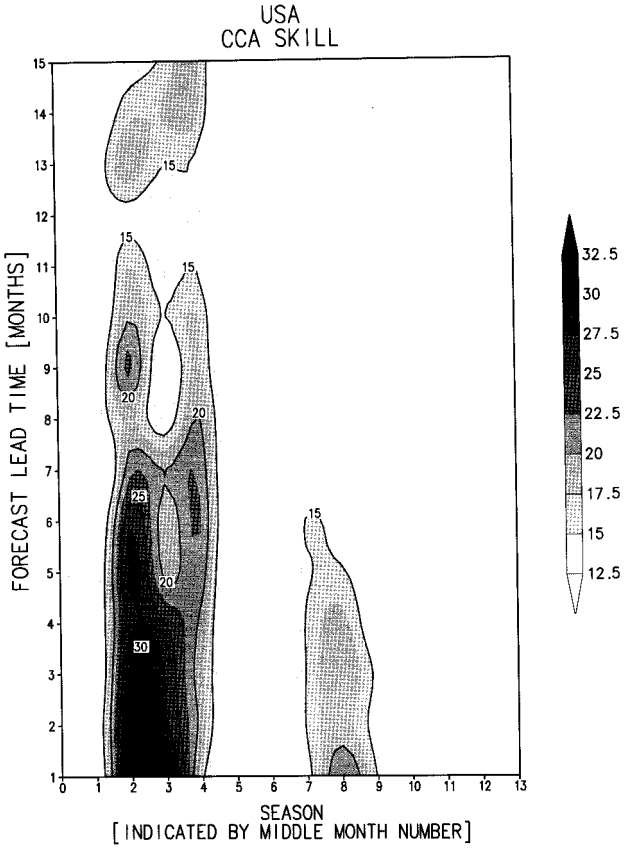

FIG. 15. As in Fig. 3 except for CCA forecasts over the United States with a predictor combination of surface air temperature, geopotential height field at 700 hPa, and quasi-global SST. The SST field is weighted double.

that found here, being overshadowed by relatively stronger relationships involving southern France and the western Mediterranean Sea in general. The associated forecast skill is found mainly in late summer and autumn, and is found to be related to the worldwide tropical SST anomaly rather than the NAO.

The above discussion points at possible ways to improve the design of the CCA methodology. Consider the target season to be in winter where we have maximum skill. The present methodology uses four contiguous 3-month predictor periods. Regardless of the forecast lead time there will always be one predictor period in the winter season, albeit not always the core winter period of JFM. As seen from Fig. 12b there is considerable longer term, interdecadal, variability of the NAOI. This would suggest that considerably more predictor periods should be used to achieve a better resolution of the status of this interdecadal variability. An alternative, similar way to obtain information on the relevant interdecadal variations could be to consider, for example, only the four winter periods, NDJ, DJF, JFM, and FMA, but doing this for several years prior to the target season. Data from the nonwinter seasons that may not contribute much to forecast skill, or worse introduce noise, could thereby be omitted.

We have seen the imprint of something akin to the NAO in the geopotential height field at 700 hPa. Upper-air data are, however, unfortunately only available from approximately the last 50 yr. However, the NAO is also discernible in the surface pressure field. The use of this field would lengthen the time series considerably and thereby give further insight into the empirical relations revealed in the present study as well as improve the assessment of skill. The work by Hurrell and van Loon (1994) give important indications of what to expect. By studying an NAO index (similar to the one used in this study) for the 130-yr time period 1864–1993, they find considerable nonstationarity in the power spectrum. The early part of the record is dominated by variance at biennial periods while the latter part is dominated by variability with periods of 6–10 yr. It is therefore possible that the empirical relationships obtained by a linear technique like CCA will prove to be useful for predictions for only that part of the record that has similar nature of variability as that during the time period for which it has been trained. This fact does not, however, imply the absence of predictability in time periods characterized by other types of variability, but rather that an empirical method such as CCA is unable to quickly

accommodate nonstationarities. There is no reason, for example, why the CCA technique should not be able to avail itself of the QBO-like behavior in the NAO found in the latter part of the 1800s if we consider only this time period. Because finely resolved field data are not readily available for this earlier period, the level of skill realizable for that period using a pattern relationship technique such as CCA remains an open question.

To obtain a more complete assessment of the potential of the NAO in seasonal forecasting, it is necessary to understand the physical mechanisms that are responsible for this phenomenon. At the present time we have some descriptive understanding of the NAO while an understanding of the physical mechanisms seems to be lacking. Such a physical understanding would provide the tools to estimate the extent to which the NAO varies at low frequencies.

It is interesting to speculate about possible physical mechanisms that are responsible for the NAO and especially its low-frequency variability. It is then natural to look into possible links to the ocean with its longer timescales. One such intriguing link is the interplay between the surface air wind conditions above, and the deep water formation in, the North Atlantic Subpolar Sea (Norwegian, Iceland, Greenland, and Barents Seas). The interior of this sea is characterized by a ''mediterranean'' type of circulation; that is, the water that enters into the basin becomes denser (in this case due to cooling) to the point that convection starts and deep water is formed (Stigebrandt 1985). However, there is more than enough freshwater supply from surrounding rivers and the relatively fresher Arctic Ocean to—given the possibility to enter into the interior of the basin—lower the density beyond what is necessary to halt the deep-water formation. The climatological wind field over the basin plays here a key role. The rather persistent cyclonic circulation around the Icelandic low keeps, by the induced Ekman transports, the fresher water in coastal currents. In the absence of this cyclonic atmospheric circulation the conditions in the subpolar sea would be similar to that in the Arctic Ocean—that is, covered by perennial sea ice. Such a drastic change is not very likely but the amplitude and timescale of the NAO is nevertheless large enough to create anticyclonic conditions several winters in a row.

Several recent studies have indicated the existence of variability on the interdecadal timescale in the North Atlantic Subpolar Sea [see, e.g., Weaver and Hughes (1992) for a review]. Here we will just mention some recent observational and modeling studies that suggest that interdecadal variability in the ocean indeed occurs and where a link to the NAO is possible. The role of freshwater supply to the northern North Atlantic has been studied in an oceanic GCM by Rahmstorf (1995). That study shows that when freshwater is added (artificially) to the northern North Atlantic it is possible to obtain variability on the interdecadal timescale. The realism of this variability and possible links to the NAO are interesting questions. Recent observations in the North Atlantic Subpolar Sea show that a freshening has occured during the period 1965–95. Blindheim et al. (1996) argue that this is caused by changed wind conditions in the area, which cause low saline water from the East Greenland Current to enter into the interior of the Subpolar Sea.

Acknowledgments. We thank Drs. Hans Alexandersson, SMHI; Hannu Savijärvi, University of Helsinki; Raino Heino, FMI; Theo Opsteegh, KNMI; and Trausti Jónsson, IMO, for helping us to get access to quality-controlled station data from their respective country. The Carbon Dioxide Information Analysis Center (CDIAC) is gratefully acknowledged for making their data available for us. Discussions with Prof. Anders Stigebrandt on possible relationships between the NAO and oceanographical processes in the Nordic Seas were very enlightening. We furthermore thank Drs. Brian Reinhold, Hans Alexandersson, and the three anonymous reviewers for their constructive criticism, which helped to improve the presentation.

REFERENCES

Barnett, T. P., and R. Preisendorfer, 1987: Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Mon. Wea. Rev.,* **115,** 1825–1850.

Barnston, A. G., 1994: Linear statistical short-term climate predictive skill in the Northern Hemisphere. *J. Climate,* **7,** 1513–1564.

——, and R. E. Livezey, 1987: Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Mon. Wea. Rev.,* **115,** 1083–1126.

——, and Coauthors, 1994: Long-lead seasonal forecasts—Where do we stand? *Bull. Amer. Meteor. Soc.,* **75,** 2097–2114.

Blindheim, J., V. Borovkov, B. Hansen, S. A. Malmberg, W. Turrell, and S. Österhus, 1996: Recent upper layer cooling and freshening in the Norwegian Sea. Contributions of Statutory Meetings, International Council for the Exploration of the Sea, 12 pp.

Branković, C., T. N. Palmer, and L. Ferranti, 1994: Predictability of seasonal atmospheric variations. *J. Climate,* **7,** 217–237.

Bretherton, C. S., C. Smith, and J. M. Wallace, 1992: An intercomparison of methods for finding coupled patterns in climate data. *J. Climate,* **5,** 541–560.

Craddock, J. M., and R. Ward, 1962: Some statistical relationships between the temperature anomalies in neighbouring months in Europe and western Siberia. Meteorological Office London, Scientific Paper 12, 31 pp. [Available from U.K. Meteorological Office, London Road, Bracknell, Berkshire RG12 2SY, United Kingdom.]

Deser, C., and M. Timlin, 1995: Extratropical air–sea interaction on weekly time scales. *Proc. Twentieth Annual Climate Diagnostics Workshop,* Seattle, WA, Climate Prediction Center and Joint Institute for the Study of the Atmosphere and Ocean, 300–303.

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

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

Glahn, H. R., 1968: Canonical correlation and its relationship to discriminant analysis and multiple regression. *J. Atmos. Sci.,* **25,** 23–31.

Hastenrath, S., 1991: *Climate Dynamics of the Tropics.* Kluwer Academic Publishers, 488 pp.

Hurrell, J. W., 1995: Decadal trends in the North Atlantic Oscillation: Regional temperatures and precipitation. *Science,* **269,** 676–679.
——, and H. van Loon, 1994: Analysis of low-frequency climate variations over the North Atlantic using historical atmospheric data. *Proc. Nineteenth Annual Climate Diagnostics Workshop,* College Park, MD, NOAA/Climate Analysis Center and Department of Meteorology, University of Maryland, 30–33.
Latif, M., T. P. Barnett, M. A. Cane, M. Flügel, N. E. Graham, H. von Storch, J. S. Xu, and S. E. Zebiak, 1994: A review of ENSO prediction studies. *Climate Dyn.,* **9,** 167–179.
Livezey, R. E., 1990: Variability of skill of long-range forecasts and implications for their use and value. *Bull. Amer. Meteor. Soc.,* **71,** 300–309.
Lorenz, E. N., 1956: Empirical orthogonal functions and statistical weather prediction. Scientific Report 1, Statistical Forecasting Project, Massachusetts Institute of Technology, Cambridge, MA, 49 pp.
——, 1969: The predictability of a flow which possesses many scales of motion. *Tellus,* **21,** 289–307.
Michaelsen, J., 1987: Cross-validation in statistical climate forecast models. *J. Climate Appl. Meteor.,* **26,** 1589–1600.
Namias, J., 1953: *Thirty-Day Forecasting: A Review of a Ten-Year Experiment. Meteor. Monogr.,* No. 33, Amer. Meteor. Soc., 279–285.
Nap, J. L., H. M. van den Dool, and J. Oerlemans, 1981: A verification of long range weather forecasts in the seventies. *Mon. Wea. Rev.,* **109,** 306–312.
Nicholls, N., 1980: Long-range weather forecasting: Value, status and prospects. *Rev. Geophys. Space Phys.,* **18,** 771–788.
Palmer, T. N., and Z. Sun, 1985: A modeling and observational study of the relationship between sea surface temperature anomalies in the northwest Atlantic and the atmospheric general circulation. *Quart. J. Roy. Meteor. Soc.,* **111,** 947–975.
Preisendorfer, R. W., 1988: *Principal Components Analysis in Meteorology and Oceanography.* Vol. 17, *Developments in Atmospheric Sciences,* Elsevier, 425 pp.
Rahmstorf, S., 1995: Bifurcations of the Atlantic thermohaline circulation in response to changes in the hydrological cycle. *Nature,* **378,** 145–149.
Reynolds, R. W., 1988: A real-time global sea surface temperature analysis. *J. Climate,* **1,** 75–86.
Rossby, C. G., 1941: The scientific basis of modern meteorology. *Climate and Man. Yearbook of Agriculture,* U.S. Govt. Printing Office, 599–655.
Slutz, R., S. J. Lubler, J. D. Hiscox, S. D. Woodruff, R. J. Jenne, D. H. Joseph, P. M. Steurer, and J. D. Elius, 1985: Comprehensive Ocean Atmosphere Data Set. National Oceanic and Atmospheric Administration, 1262 pp. [Available from Climate Research Program, ERL, R/E/AR6, 325 Broadway, Boulder, CO 80303.]
Stigebrandt, A., 1985: On the hydrographic and ice conditions in the northern North Atlantic during different phases of a glaciation cycle. *Palaeogeogr., Palaeoclimatol., Palaeoecol.,* **50,** 303–321.
Trenberth, K. E., and J. W. Hurrell, 1994: Decadal atmosphere–ocean variations in the Pacific. *Climate Dyn.,* **9,** 303–319.
Van den Dool, H. M., 1984: Long-lived air temperature anomalies in the midlatitudes forced by the surface. *Mon. Wea. Rev.,* **112,** 555–562.
——, 1994: Long-range weather forecasts through numerical and empirical methods. *Dyn. Atmos. Oceans,* **20,** 247–270.
——, and J. L. Nap, 1981: An explanation of persistence in monthly mean temperatures in the Netherlands. *Tellus,* **33,** 123–131.
——, and R. E. Livezey, 1984: Geographical distribution and seasonality of month-to-month correlation of monthly mean 700 mb heights. *Mon. Wea. Rev.,* **112,** 610–615
——, and J. L. Nap, 1985: Short and long range air temperature forecasts near an ocean. *Mon. Wea. Rev.,* **113,** 878–886.
——, and S. Saha, 1990: Frequency dependence in forecast skill. *Mon. Wea. Rev.,* **118,** 128–137.
van Loon, H., and J. C. Rogers, 1978: The seesaw in winter temperatures between Greenland and northern Europe. Part I: General description. *Mon. Wea. Rev.,* **106,** 296–310.
Vedin, H., H. Alexandersson, and M. Persson, 1991: Use of persistence in temperature and precipitation for forecasts of spring floods. SMHI Meteorology, 82, Swedish Meteorological and Hydrological Institute, Norrköping, Sweden, 33pp.
Walker, G. T., and E. W. Bliss, 1932: World weather. V. *Mem. R. Meteor. Soc.,* **4,** 53–84.
Wallace, J. M., and D. S. Gutzler, 1981: Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Mon. Wea. Rev.,* **109,** 784–812.
Weaver, A. J., and T. M. C. Hughes, 1992: Stability and variability of the thermohaline circulation and its links to climate. *Trends Oceanogr.,* **1,** 15–70.