Forecast Verification of the Polar Ice Prediction System (PIPS) Sea Ice Concentration Fields*

MICHAEL L. VAN WOERT+ AND CHENG-ZHI ZOU+

NOAA/NESDIS/Office of Research and Applications, Camp Springs, Maryland

WALTER N. MEIER#

United States Naval Academy, Annapolis, Maryland

PHILIP D. HOVEY+

NOAA/NESDIS/Office of Research and Applications, Camp Springs, Maryland

RUTH H. PRELLER AND PAMELA G. POSEY

Naval Research Laboratory, Stennis Space Center, Mississippi

(Manuscript received 26 February 2003, in final form 30 October 2003)

ABSTRACT

The National Ice Center relies upon a coupled ice–ocean model called the Polar Ice Prediction System (PIPS) to provide guidance for its 24–120-h sea ice forecasts. Here forecast skill assessments of the sea ice concentration (C) fields from PIPS for the period 1 May 2000–31 May 2002 are presented. Methods of measuring the sea ice forecast skill are adapted from the meteorological literature and applied to locations where the forecast or analysis sea ice fields changed by at least ±5%. The forecast skill referenced to climatology was high (>0.85, relative to a maximum score of 1.0) for all months examined. This is because interannual variability in the climatology, which is used as a reference field, is much greater than the day-to-day variability in the forecast field. The PIPS forecasts were also evaluated against persistence and combined climatological–persistence forecasts. Compared to persistence, the 24-h forecast was found to be skillful (>0.2) for all months studied except during the freeze-up months of December 2000 and January 2001. Relative to the combined reference field, the 24-h forecast was also positive for the non-freeze-up months; however, the skill scores were lower ($\sim$0.1). During the poorly performing freeze-up months, a linear combination of persistence ($\sim$95% weight) and climatology ($\sim$5% weight) appears to provide the best available sea ice forecast.

To examine the less restrictive question of whether PIPS can forecast sea ice concentration changes, independent of the magnitude of the changes, ''threat indexes'' patterned after methods developed for tornado forecasting were established. Two specific questions were addressed with this technique. The first question is: What is the skill of forecasting locations at which a *decrease* in sea ice concentration has occurred? The second question is: Does PIPS correctly forecast *melt-out* regions? Using the more relaxed criterion of a threat index for defining correct forecasts, it was found that PIPS correctly made 24-h forecasts of decreasing sea ice concentration $\sim$10%–15% of the time (it also correctly forecast *increasing* sea ice concentration an additional $\sim$10%–15% of the time). However, PIPS correctly forecast melt-out conditions <5% of the time, suggesting that there may be deficiencies in the PIPS parameterization of marginal ice zone processes and/or uncertainties in the atmospheric–oceanic fields that force PIPS.

1. Introduction

The meteorological community has a long-standing history of environmental forecasting. In concert with developing new forecast systems, that community has put considerable energy into developing methods for assessing changes in forecast skill. Operational ocean forecasting systems are in their infancy, but there is considerable interest in developing robust, skillful forecast systems for that environment (e.g., Koblinsky and Smith 2001; Pinardi and Woods 2002).

[*] Naval Research Laboratory Contribution Number NRL/JA/7320/03/122.

[+] Additional affiliation: National Ice Center, Suitland, Maryland.

[#] Current affiliation: National Snow and Ice Data Center, Boulder, Colorado.

Corresponding author address: Dr. Michael Van Woert, NOAA/NESDIS/OSDPD/National Ice Center, E/SP, Federal Office Bldg. #4, Rm. 1069, 5200 Auth Rd., Camp Springs, MD 20746-4304.
E-mail: mvanwoert@natice.noaa.gov

Sea ice gained recognition by the U.S. Navy as a hazard to navigation when it caused severe damage to a convoy of ships navigating along the west coast of Greenland during the establishment of a distant early warning station and Thule Air Force Base, Greenland. In response to this situation, the navy established the Naval Ice Center with a mandate to chart and forecast global sea ice conditions. During the 1970s the National Oceanic and Atmospheric Administration (NOAA) was formed and joined forces with the navy, establishing the Joint Ice Center (JIC). In 1995 the U.S. Coast Guard pooled its resources with the navy and NOAA personnel to establish the National/Naval Ice Center (NIC), the sole U.S. center for operational sea ice analysis and forecasting.

Currently the NIC uses the Polar Ice Prediction System (PIPS) version 2.0 as the basis for its ‘‘operational’’ short-term (24–120 h) sea ice forecasts. These forecasts are evaluated daily and amended by skilled analysts using reconnaissance data (if available), the most recent weather charts and data, and historical knowledge of the conditions in the area to provide the highest quality forecasts possible out to 120 h. Special emphasis in these forecasts is placed on the location of the ice edge and the conditions in the marginal ice zone (MIZ), as these are the most critical operational areas for marine transportation and safety.

Here we focus attention on evaluating the sea ice concentration (C) forecast fields from PIPS for the 25-month period 1 May 2000–31 May 2002. The goal of this study is not to denigrate PIPS, as it is one of the few examples of an ocean forecast system that is actually used operationally. Rather, the broader goal is to illustrate the importance and some of the basic issues involved in assessing the skill of ocean forecast systems. Methods of assessing the sea ice forecast skill are adapted from well-established methodologies developed by the meteorological community. In the next section an overview of PIPS is provided. Section 3 introduces the methodologies used for assessing forecast skill and discusses differences adopted for the sea ice forecasting problem. Section 4 describes the results of this study and section 5 concludes with a discussion and summary of the results.

2. The Polar Ice Prediction System

PIPS was developed at the Naval Research Laboratory's (NRL) Stennis Space Center (Preller 1992, 1999; Cheng and Preller 1996; Preller and Posey 1989) and runs operationally at the U.S. Fleet Numerical Meteorology and Oceanography Center (FNMOC) in Monterey, California. PIPS is a fully coupled ice–ocean model forced by atmospheric forecast products. The oceanic model is a variant of the Cox (1984) model (Cheng and Preller 1996) with the ETOP05 bottom bathymetry (Heirtzler 1985). It is constrained to the Levitus (1982) climatology using a time constant of 250 days. The ocean model is coupled to a dynamic–thermodynamic sea ice model, which incorporates a viscous–plastic constitutive law (Hibler 1979, 1980). It uses a two-level ice thickness scheme, with the ‘‘thick ice’’ further subdivided into seven subcategories (Walsh et al. 1985). Atmospheric forcing is provided by the Navy Operational Global Atmospheric Prediction System (NOGAPS), which provides weather forecasts out to 120 h (Hogan and Rosmond 1991). The grid resolution of PIPS is 0.28°, which varies from 17 to 33 km depending upon the location of the grid square within the spherical coordinate system. The final output is converted to a fixed 18 km × 18 km grid (Fig. 1). To facilitate comparison with other published products, all PIPS forecast and observed fields have been interpolated to the 25-km-resolution National Snow and Ice Data Center polar stereographic projection tangent at 70° (Weaver et al. 1987).

With the exception of sea ice concentration and thickness, all other variables are initialized from the previous day's 24-h forecast. The sea ice concentration field in the PIPS forecast system is initialized daily ($T = 0$ h) using the Cal/Val algorithm (Hollinger 1991), which is based on Special Sensor Microwave Imager (SSM/I) data (Posey and Preller 1994; see also Preller et al. 1992). In the initialization scheme, PIPS ice concentrations at locations where the observed Cal/Val ice concentrations are $<50\%$ or $>80\%$ and where the differences between the PIPS and Cal/Val concentrations are $>5\%$ and $>10\%$, respectively, are replaced with the Cal/Val values. Since the PIPS ice concentrations are $\sim 95\%$ or greater in most places and the Cal/Val values typically saturate to 100% except near the ice edge (Partington 2000; Meier et al. 2001), the model uses the Cal/Val data mostly near the ice edge. Moreover, as discussed more completely below, neither PIPS nor the Cal/Val algorithm easily differentiates between open-water areas and thin ice concentrations; thus, the 15% ice concentration isopleth is taken as the ice edge.

Sea ice thickness is initialized from the previous forecast, but the ice edge is modified in the following way: if the SSM/I concentration indicates that no ice is present at a location where the model predicted ice, the ice is removed and the temperature of the mixed layer is raised to 1°C above freezing. In contrast, if the SSM/I concentration indicates that ice is present, but the model did not predict ice, the mixed layer temperature is set to freezing and an ice-concentration-dependent ice thickness is added. In particular, if the ice concentration is less than 50%, the ice thickness is set to 0.5 m and if it is greater than 50%, it is set to 1 m. In the event that the model must be reinitialized, it is restarted from climatology as described by Preller and Posey (1989).

Regions with missing data, such as near the pole, are estimated by optimal interpolation from nearby points, thus providing complete, daily, hemispheric analyses. The valid range for both the initialization and forecast fields is 0%–100% representing the range from ice-free

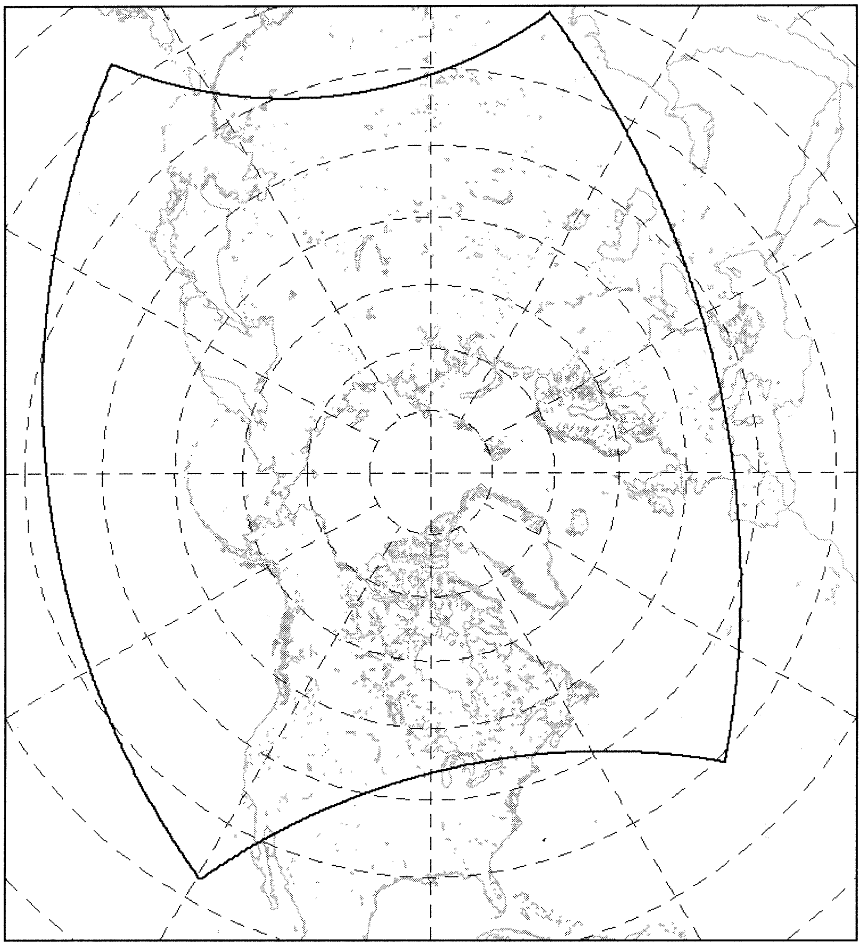

FIG. 1. Spatial domain of PIPS. The latitude spacing is 10° and the longitude spacing is 30°.

conditions to complete ice cover. (Note: ice concentration, or fractional ice cover, and the statistics that are derived from it are unitless.)

3. Methods of forecast skill assessment

a. Skill scores

Forecasting sea ice drift and the ice edge are analogous to the problem of forecasting wind velocity and the location of weather fronts in atmospheric models. Similarly, sea ice thickness and concentration are scalar fields analogous to atmospheric temperature in weather models. Some work has been done to verify ice drift models (Flato and Hibler 1992; Grumbine 1998) and ice edge prediction (Pritchard et al. 1990), but with the exception of the studies by Preller and Posey (1996) and Van Woert et al. (2001), little work has been done to evaluate operational sea ice concentration forecasts.

To assess the forecast skill of PIPS, the forecast ice concentration changes were compared to ‘‘truth,’’ which, for consistency, is taken to be the PIPS sea ice concentration analysis at the valid time of the forecast. This satellite-derived sea ice analysis is less than perfect. However, by using a common analysis to both initialize the model and serve as truth, skill scores allow the model performance to be evaluated in a mode that is relatively insensitive to observational errors and deficiencies in the structure of the observing system.

Forecast skill is typically defined in terms of a generic measure of accuracy, A, as follows:

$$\mathrm{SS} = \frac{A_f - A_r}{A_p - A_r}, \tag{1}$$

where A_f, A_p, and A_r denote the accuracy of the forecast system, the accuracy of a perfect forecast, and the accuracy of a reference forecast, respectively (Brier and Allen 1951; Murphy and Daan 1985). In this formulation the skill score represents the improvement in accuracy of a forecast with respect to an as yet to be defined reference forecast, relative to the total improvement in accuracy.

Murphy (1988) and Murphy and Epstein (1989) advocate the use of the mean-square error (MSE), defined as

$$\mathrm{MSE}(a, b) = N^{-1} \sum_i (a_i - b_i)^2 \qquad (i = 1, \ldots, N), \tag{2}$$

as the measure of accuracy. Substitution of (2) into (1) gives

$$\mathrm{SS} = \frac{\mathrm{MSE}(f, O) - \mathrm{MSE}(R, O)}{\mathrm{MSE}(P, O) - \mathrm{MSE}(R, O)}, \tag{3}$$

where f and P represent the system forecast and perfect forecast fields, respectively, and O and R represent the analyzed (observed) and reference fields at the valid forecast time. Since the MSE for a perfect forecast, MSE(P, O), is zero, Eq. (3) can be written as

$$\mathrm{SS} = 1 - \frac{\mathrm{MSE}(f, O)}{\mathrm{MSE}(R, O)}. \tag{4}$$

From (4) it is seen that SS is positive (negative) when the MSE for the reference field is greater (less) than the MSE for the forecast field. Moreover, for a perfect forecast [$\mathrm{MSE}(f, O) = 0$] $\mathrm{SS} = 1$, and for no forecast skill [$\mathrm{MSE}(f, O) \geq \mathrm{MSE}(R, O)$] $\mathrm{SS} \leq 0$.

Murphy (1988) showed that the square of the *correlation coefficient* [also commonly referred to as the anomaly correlation coefficient (Brier and Allen 1951)], another widely used measure of forecast skill (Arpe et al. 1985), can be viewed as the *potential* forecast skill (i.e., the maximum attainable SS when all biases are eliminated). For weather forecasting a correlation coefficient greater than an arbitrary value of 0.6 is normally considered to be a ''skillful'' forecast (Hollingsworth et al. 1980). This corresponds to a skill score of ~0.36. It is clear from (4), however, that any skill score that is greater than zero represents an improvement over the reference forecast.

A critical element of forecast verification is the selection of the reference value, or the zero point on the scale on which skill is measured. In weather forecasting, persistence is often taken as the appropriate reference for measuring the skill of short-range forecasts and climatology is frequently used for medium- and long-range forecasts. The skill score relative to climatology, SS_c, based on (4) is given by

$$\mathrm{SS}_c(n) = 1 - \frac{\mathrm{MSE}(f_n, O_n)}{\mathrm{MSE}(C_n, \mathrm{O}_n)} \tag{5}$$

(Murphy and Epstein 1989; Murphy 1988), where f_n is the n-h forecast field (n = 24, 48, 72, and 120 h), and O_n and C_n are the analyzed and climatological fields at the valid time for the forecast, respectively.

Murphy (1988) discusses the nuances of using various definitions of climatology. In particular he showed that, by virtue of its improvement in defining the reference state, in most cases the use of a multiple-valued climatology (e.g., daily climatology) should give a lower skill score than a single-valued climatology (e.g., annual climatology). In this study a daily sea ice concentration climatology was developed by averaging the daily output from the Cal/Val sea ice concentration algorithm for the period 1 June 1995–31 May 2002.

For short-term forecasts it is customary in the meteorological field to use persistence as the reference field. (A persistence forecast is obtained by assuming that the analyzed sea ice conditions on a given day are the forecast conditions, i.e., conditions will stay the same.) PIPS is initialized daily with observed sea ice fields and over time scales of a few days or more is observed to move predominately in response to the imposed wind forcing. Thus, it is reasonable to consider persistence as a reference state for short-term sea ice forecast evaluation. Under this assumption, (4) becomes

$$\mathrm{SS}_p(n) = 1 - \frac{\mathrm{MSE}(f_n, O_n)}{\mathrm{MSE}(O_0, O_n)}, \tag{6}$$

where $\mathrm{SS}_p(n)$ represents the skill score relative to persistence, O_0 is the analyzed field at the time the forecast was made, and all other parameters have previously been defined.

Under the normally reasonable assumptions that the means and variances at $T = 0$ and $T = n$ (n = 24, 48, 72, and 120 h) are equal, respectively, Murphy (1992) showed that the best choice for the reference field, R_n, is a linear combination of climatology and persistence (CLIPER), given by

$$R_n = r(n)O_0 + [1 - r(n)]C_n, \tag{7}$$

where $r(n)$ is the first-order autocorrelation coefficient between the observed values at $T = 0$ and at $T = n$ (n = 24, 48, 72, and 120 h) given by

$$r(n) = \frac{\sum (O_0 - \overline{O}_0)(O_n - \overline{O}_n)}{\sqrt{\sum (O_0 - \overline{O}_0)^2 \sum (O_n - \overline{O}_n)^2}}. \tag{8}$$

For the monthly statistics discussed in this study, the sums in (8) are taken over all available space and time points for the month. For $r(n) \cong 1$ (i.e., the observations on any given day are highly correlated with future conditions on time scales of a few days), the persistence term dominates R_n. In contrast, when $r(n) \cong 0$, R_n is dominated almost exclusively by climatology. Given this definition of R_n, the skill score relative to CLIPER, $\mathrm{SS}_{\mathrm{cp}}(n)$, is defined as

$$\mathrm{SS}_{\mathrm{cp}}(n) = 1 - \frac{\mathrm{MSE}(f_n, O_n)}{\mathrm{MSE}(R_n, O_n)}. \tag{9}$$

In principle, this methodology can be applied directly to the ice forecasting problem presented here. However, in practice, much of the ice concentration field (in particular the high Arctic and open-ocean regions) remains persistently unchanged. Since the model correctly predicts these large unchanged regions, the resulting skill scores would be impressively high (Van Woert et al. 2001). The goal of a skill score, however, is to evaluate a model's ability to forecast significant geophysical change. The accuracy of the SSM/I-derived ice concentrations is poorly characterized in the marginal ice zone, but is believed to be ~5%–10% (Steffen et al. 1992). Thus, to accurately assess the ability of PIPS to forecast sea ice concentration change, *only grid nodes for which the forecast or observed fields changed by a specified amount relative to the previous day's estimates are included in the estimation of the skill scores.*

Based on the results of the Steffen et al. (1992) study, ±5% was selected as the primary threshold for change. However, the CLIPER skill scores were also computed using a ±15% cutoff to examine the sensitivity of the results to the specific choice of threshold. For the ±5% threshold ~2500 grid nodes per day out of the available ~16 000 ice-covered grid nodes were included in the analysis. All totaled, ~75 000 grid nodes per month were used to compute these statistics. Ninety-five percent confidence limits on the skill scores were estimated using the nonparametric, bootstrap technique (Efron 1979a,b, Press et al. 1992). The bootstrap method makes no assumptions about the probability distribution function. Uncertainties are estimated by selecting N data values with replacement from the N original data points and computing the desired parameter. This process is repeated many times (in this case 1000 times). The best estimate of the parameter is the median of the 1000 parameter estimates. The upper and lower 95% confidence limits are obtained from the upper and lower 2.5 percentiles of the distribution.

b. Threat indexes

An important function of a forecast system is to warn users of substantial changes in an environmental parameter. For sea ice, these ''warnings'' are particularly important, because the impact of slow changes can often be mitigated while large sudden changes can be extremely dangerous, putting life and property at risk. One method of assessing a model's ability to forecast rare, but sudden, large changes is through the use of a ''threat index'' (Ghil et al. 1979; Atlas et al. 1981; Murphy and Daan 1985). These statistics have been used previously for evaluating the forecast accuracy of severe weather events such as tornadoes (Stephenson 2000) and for validating cloud parameterizations of numerical weather prediction models (Mace et al. 1998; Beesley et al. 2000). In this methodology, the model forecasts an event to ''occur'' or ''not occur'' and the outcome is either ''correct'' or ''incorrect.'' The results of this test are represented by a 2 × 2 contingency table. When a forecast correctly predicts an event to occur, it is termed a hit. When a forecast correctly predicts an event to not occur, it is termed a correct rejection. Hits and correct rejections are both correct forecasts. When an event occurs but the forecast fails to predict its occurrence, it is termed a miss, and finally, when an event does not occur, but the forecast predicts that it should, it is termed a false alarm. Misses and false alarms are both incorrect forecasts.

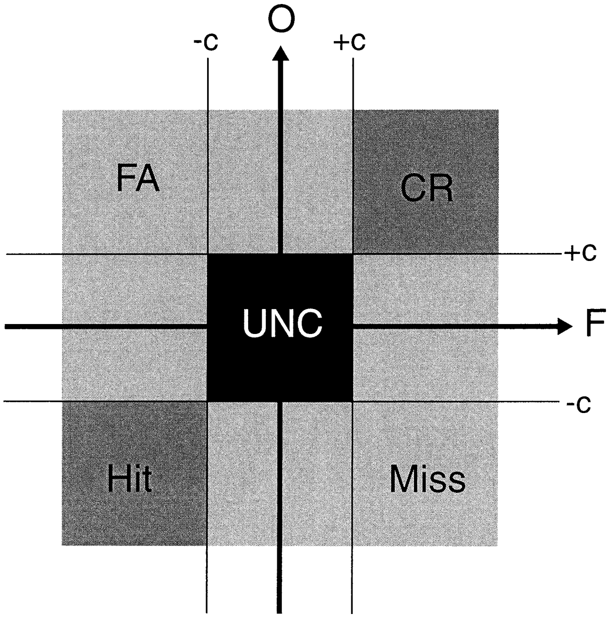

FIG. 2. Contingency diagram for threat indexes: F and O represent the forecast and observed values, respectively; hit, miss, FA, CR, and UNC represent hits, misses, false alarms, correct rejections, and unchanged, respectively. All are defined in the text. Here, ±c represents the positive and negative cutoff values (±5% or ±15%) used in this study.

In the context of sea ice forecasting, the threat index is used here to address two specific questions. The first is: What is the skill of forecasting locations at which a geophysically meaningful *decrease* in ice concentration has occurred? This is a useful question for ship navigation, because it provides an indication of whether the ice is opening, which for a ship beset by ice, might mean the difference between needing to mobilize or cancel a rescue attempt. As was the case for the skill scores, a threshold of ±5% was applied to the data to determine locations where ''significant'' change occurred. The exclusion of data that falls into the unchanged category (Fig. 2) represents a departure from the typical usage of a threat index, but again, it is necessary to avoid biasing the results with large open-water and ice-covered areas. These ''unchanged'' data are denoted for clarity in the contingency table (Fig. 2) but are not used in any subsequent statistical calculations.

In this formulation a model forecast of a ≥5% decrease in ice concentration when a ≥5% decrease actually occurs is considered a hit. It is important to note, however, that this skill score is only testing for substantial decrease in the ice cover (with a threshold of ±5%). It does not assess how accurate the decrease is. For example, a forecast decrease in ice cover of 6% when the actual decrease was 90% is still counted as a correct forecast. In addition, when a ≥5% increase in sea ice concentration is both predicted and observed, it is again a correct forecast (correct rejection), but may not be of operational significance (except perhaps to expedite the rescue effort). Similarly, when a ≥5% decrease in sea ice concentration is predicted, but not observed, the forecast is considered a false alarm and if the model fails to predict an observed ≥5% decrease in ice cover it is termed a miss.

The second question is: Does PIPS correctly forecast regions of *melt out* (a hit) or conversely *freeze up* (a correct rejection)? Highly accurate forecasts of this nature are particularly useful to ships operating at the ice edge or in the MIZ that need to know whether there will be a sudden appearance of sea ice that could endanger a ship. These situations are well suited to a threat index analysis, because the surface in these regions can generally be viewed as either being ice covered or ice free at any given time. However, application of this methodology to MIZ forecast assessment, while straightforward, requires some modification from the standard meteorological treatment because neither the model nor the remotely sensed ice concentration products precisely discriminate the ice edge.

In particular, for the SSM/I data, the relatively low spatial resolution of the channels used in the algorithms (~25 km) limits the precision of the ice edge detection. Moreover, the ice concentration algorithms often have difficulty discriminating thin ice from open water because they are usually ‘‘tuned’’ for higher-concentration pack ice. Generally, the 15% ice concentration isopleth is taken to be the threshold between ice-covered and ice-free areas (Steffen et al. 1992). Similarly, PIPS does not discriminate well between ice-covered and ice-free regions at very low concentrations. Therefore, a similar threshold must be set to differentiate between model forecasts of ice-free and ice-covered areas. Cognizant of these data limitations, the threat index is established in this study by testing whether the model correctly forecasts regions of ‘‘melt out’’ ($O_0 > 15\%$ and $O_{24} < 15\%$) and ‘‘freeze up’’ ($O_0 < 15\%$ and $O_{24} > 15\%$). In contrast with the skill score analysis, only grid locations for which the forecast or observed fields changed by ±15% relative to the previous day’s values are included in the statistics.

4. Results

a. Skill scores

The monthly and spatially averaged skill scores relative to climatology [Eq. (4)] at $T = 24$, 48, 72, and 120 h are shown in Fig. 3a. The skill scores relative to climatology at $T = 24$ h are ~0.85. Examination of the MSE statistics that compose $\mathrm{SS}_c(24)$ indicates that the exceptionally high skill scores at $T = 24$ h are due to the highly variable nature of climatology [$\mathrm{MSE}(C_{24}, O_{24}) \sim 900$] relative to forecast variability [$\mathrm{MSE}(f_{24}, O_{24}) \sim 150$]. That is, the year-to-year variability at a given location greatly exceeds the forecasted day-to-day variability. In contrast, at $T = 120$ h, $\mathrm{SS}_c(120)$ has dropped to ~0.4 with considerable variability on monthly time

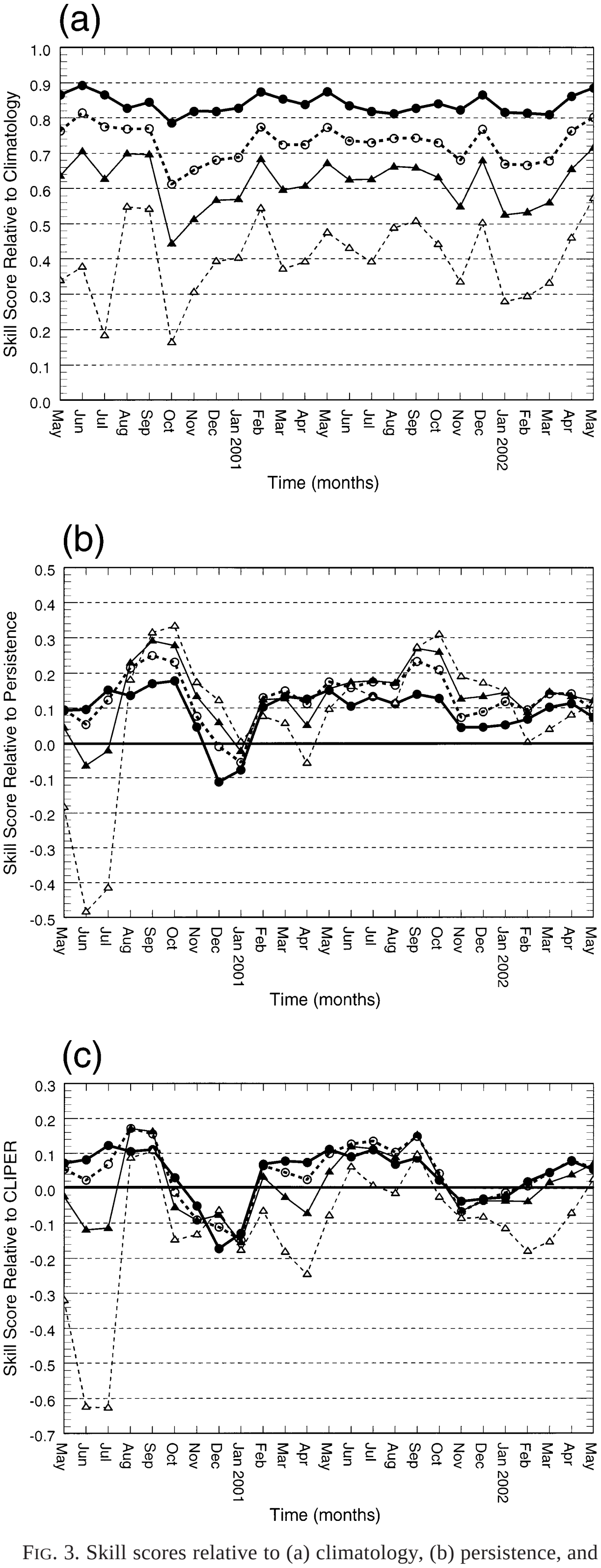

FIG. 3. Skill scores relative to (a) climatology, (b) persistence, and (c) CLIPER for the period May 2000–May 2002 based on a ±5% cutoff value. The bold solid and dashed lines with circles denote the 24- and 48-h skill scores, respectively. The solid and dashed lines with triangles denote the 72- and 120-h skill scores, respectively. The 95% confidence limits are roughly equal to the size of the dots.

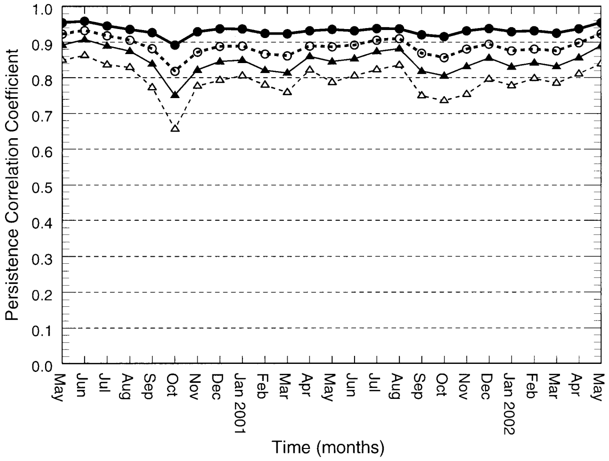

FIG. 4. Persistent correlation coefficient for the period May 2000–May 2002 based on a ±5% cutoff value. The bold solid and dashed lines with circles denote the 24- and 48-h skill scores, respectively. The solid and dashed lines with triangles denote the 72- and 120-h skill scores, respectively. The 95% confidence limits are roughly equal to the size of the dots.

scales. This value is roughly equivalent to a correlation coefficient of 0.6, which is consistent with the commonly accepted minimum correlation associated with a good weather forecast (Hollingsworth et al. 1980). On the basis of these statistics, it can be concluded that out to $T = 120$ h, PIPS performs substantially better than a climatological forecast.

The monthly and spatially averaged skill scores relative to persistence [Eq. (5)] at $T = 24$, 48, 72, and 120 h are shown in Fig. 3b. In contrast with $\mathrm{SS}_c(n)$, $\mathrm{SS}_p(n)$ hovers around 0.2 (equivalent to a correlation of ~0.45). The skill scores exhibit considerable variability on monthly time scales. Skill scores at $T = 24$ h are generally highest during spring, summer, and fall and lowest during the winter months of December and January. In particular, during December 2000 and January 2001, the skill scores turn negative indicating nonskillful 24-h forecasts. Skill scores at $T = 48$, 72, and 120 h exhibit the same general structure as the 24-h skill scores, but, at $T = 72$ and 120 h, other isolated occurrences of nonskillful forecasts also occur. It is also noteworthy that the $T = 24$ and $T = 48$ h skill scores for May 2000 are slightly greater than zero, consistent with the finding of Van Woert et al. (2001) who showed that PIPS performed better than persistence during that month. It is concluded, therefore, that, when referenced against persistence, PIPS provides a small, but significant, improvement during most months of the year.

The monthly and spatially averaged first-order autocorrelation values, $r(n)$, for lags of 24, 48, 72, and 120 h are shown in Fig 4. It is seen that $r(24)$ hovers near 0.95 for all months with little variability on monthly time scales. The only months with any detectable change in $r(24)$ were the fall months of October 2000 and 2001 with values of $r(24)$ of ~90%. This finding is also observed at the other forecast times and is noted most strongly in the $T = 120$ h value. The consistently high $r(24)$ values indicate that ‘‘persistence’’ dominates the reference field for the 24-h forecast. However, by $T = 120$ h, $r(120)$ has dropped to ~0.80 giving an ~20% weight to the climatological term in (7).

The monthly and spatially averaged CLIPER skill scores [Eq. (9)] at $T = 24$, 48, 72, and 120 h are shown in Fig. 3c. Like $\mathrm{SS}_p(n)$, $\mathrm{SS}_{\mathrm{cp}}(n)$ is mostly positive but is in general less than SS_p. The close correspondence between $\mathrm{SS}_p(n)$ and $\mathrm{SS}_{\mathrm{cp}}(n)$ and the reduction in $\mathrm{SS}_{\mathrm{cp}}(n)$ relative to $\mathrm{SS}_p(n)$ are both consistent with the analysis of Murphy (1992) who showed that when $r(n)$ is $O(1)$, CLIPER is dominated by the persistence term but gives an overall reduction in forecast skill relative to either climatology or persistence alone. It is important to note that as small as these improvements in the reference forecast are, they are large enough to expand slightly the period of nonskillful forecasts to include the months of November 2000, November–December 2001, and January 2002. Skill scores at the other forecast times generally track the $T = 24$ h results but are smaller.

To better understand the nature of the negative forecast skill scores observed during the winter months, the logarithmic, normalized histogram of the absolute differences between f_{24} and O_0 for December 2000, May and December 2001, and May 2002 were estimated (Fig. 5). It is clearly seen that during December 2000 there was a greater fraction of ‘‘poor’’ forecasts than during December 2001 [defined here as the number of points in bins with $\mathrm{ABS}(f_{24} - O_0) > 30\%$; Fig. 5a]. In addition, there was also a much greater fraction of ‘‘bust’’ forecasts [defined here as $\mathrm{ABS}(f_{24} - O_0) \sim 100\%$, the maximum possible difference]. By virtue of the squaring process used to compute it, the MSE is particularly sensitive to extreme departures from the mean. Thus, it is likely that it is the large number of bust forecasts that is responsible for the negative $\mathrm{SS}_{\mathrm{cp}}(24)$ values observed during December 2000.

Further support for this hypothesis is provided by a comparison of the May and December histograms. Overall there were many fewer bust forecasts during May than during December, consistent with the overall higher skill scores observed during May (Fig. 3c). In addition, for May 2001 there were slightly more forecast differences in the range $20 < \mathrm{ABS}(f_{24} - O_0) < 70$ than during May 2002. However, there was a slight increase in the number of large forecast differences $[\mathrm{ABS}(f_{24} - O_0) > 70]$ during May 2002. Since $\mathrm{SS}_{\mathrm{cp}}(24)$ (May 2002) $< \mathrm{SS}_{\mathrm{cp}}(24)$ (May 2001), we conclude again that it must be the large number of bust forecasts that is responsible for the observed reduction in forecast skill during some months.

None of the analysis to this point provides any information on the locations where the differences occur. To better understand the spatial distribution of the forecast skill, the monthly, $T = 24$ h CLIPER skill scores, $\mathrm{SS}_{\mathrm{cp}}(x, y)$, were estimated from (9) for each month and

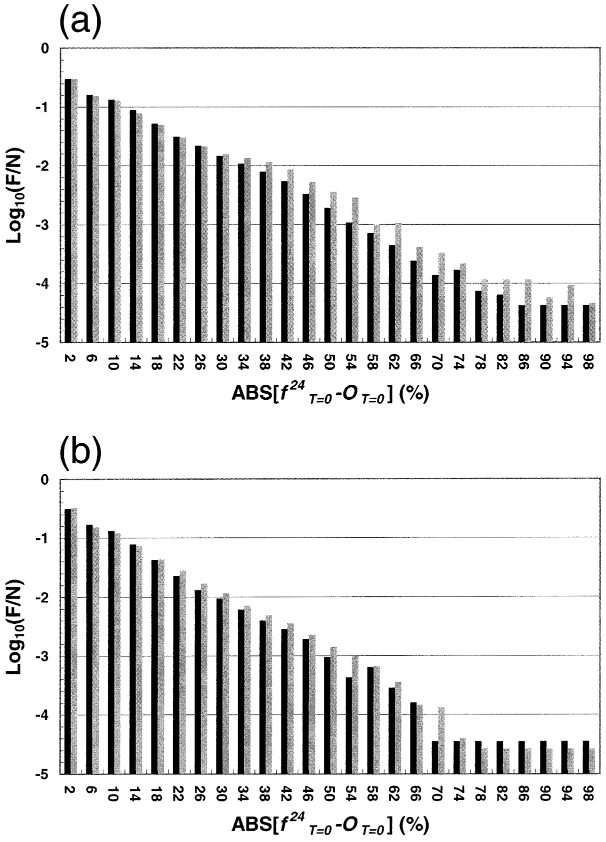

FIG. 5. Logarithmic frequency of occurrence normalized by the total number of points (N) vs the absolute difference between the 24-h forecast and observed values for (a) Dec 2000/2001 and (b) May 2001/2002. Both Dec 2000 and May 2001 are shown in gray; Dec 2001 and May 2002 are shown in black. The bins are 4% wide and the listed abscissa values denote the centers of the bins.

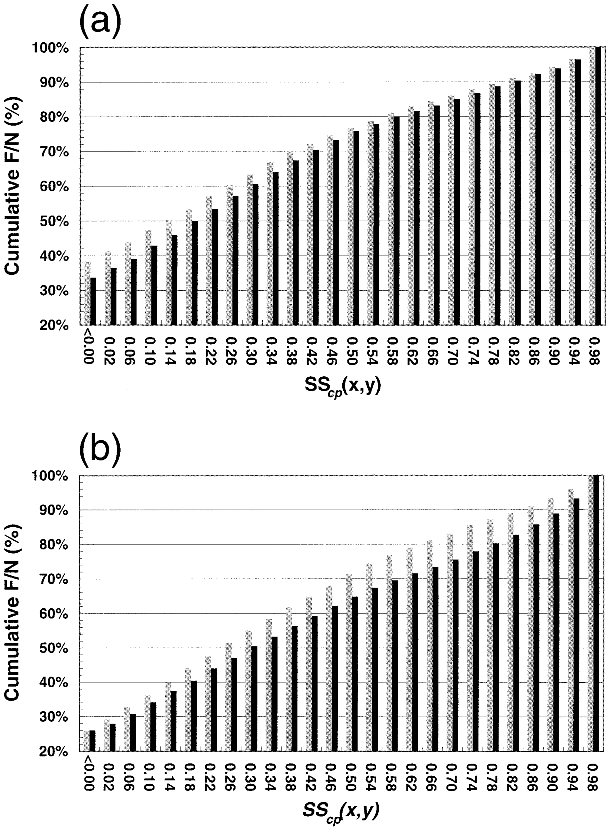

FIG. 6. Cumulative frequency of occurrence normalized by the total number of points (N) vs the skill score relative to CLIPER for (a) Dec 2000/2001 and (b) May 2001/2002. Both Dec 2000 and May 2001 are shown in gray; Dec 2001 and May 2002 are shown in black. With the exception of the first bin, the bins are 0.04 wide and the listed abscissa values denote the centers of the bins. All skill scores <0 are included in the first bin.

x, y location of the PIPS coverage. Given the monthly granularity in this statistic, the maximum possible number of points in any x, y calculation is 31. In this study we have chosen to consider only locations where changes occurred on 5 or more days to provide stability to the calculations. The threat index (next section) addresses the issue of forecast skill for isolated changes (i.e., $N < 5$).

Based on cumulative histograms of $\mathrm{SS}_{\mathrm{cp}}(x, y)$ (Fig. 6), it is seen that approximately 40% of all forecasted locations during December 2000 were nonskillful [defined here as $\mathrm{SS}_{\mathrm{cp}}(x, y) < 0.04$]. Forecast skill was slightly better during December 2001 (~35% nonskillful forecasts), consistent with Figs. 3c and 5a. However, during both Decembers studied, ~70% of all skill scores were less than 0.4. In contrast, May 2001 and 2002 had fewer (~25%) nonskillful forecasts and only ~60% of the forecasts had a skill of less than 0.4. The large number of locations with nonskillful forecasts, particularly during the winter months, suggests that the large differences observed between the forecast and observed fields (Fig. 5) affect a significant fraction of the forecasted locations at some time during the month.

The spatial distributions of $\mathrm{SS}_{\mathrm{cp}}(x, y)$ for December 2000, May and December 2001, and May 2002 are shown in Fig. 7. During May and December, and in fact all other months of study, the MIZ was the most active region of change in the PIPS domain. From the figure it is clearly seen that the MIZ undergoes a strong seasonal migration in location as well as expanding from ~3000 locations during May to ~6000 locations during December. Interannual differences also exist in the spatial distribution of poor forecasts [$\mathrm{SS}_{\mathrm{cp}}(x, y) < 0.1$; denoted by red]. As shown before, differences in $\mathrm{SS}_{\mathrm{cp}}(n)$ are strongly modulated by changes in the relative number of bust forecasts (Fig. 5). Thus, it is most likely that the poor forecast skill is driven by uncertainties in the MIZ parameterization of the model and/or inadequacies in the MIZ forcing.

The preceding statistics were based on forecasts in which the observed or forecasted fields changed by ±5%. To examine the skill of forecasting larger changes, statistics were computed based on a ±15% threshold.

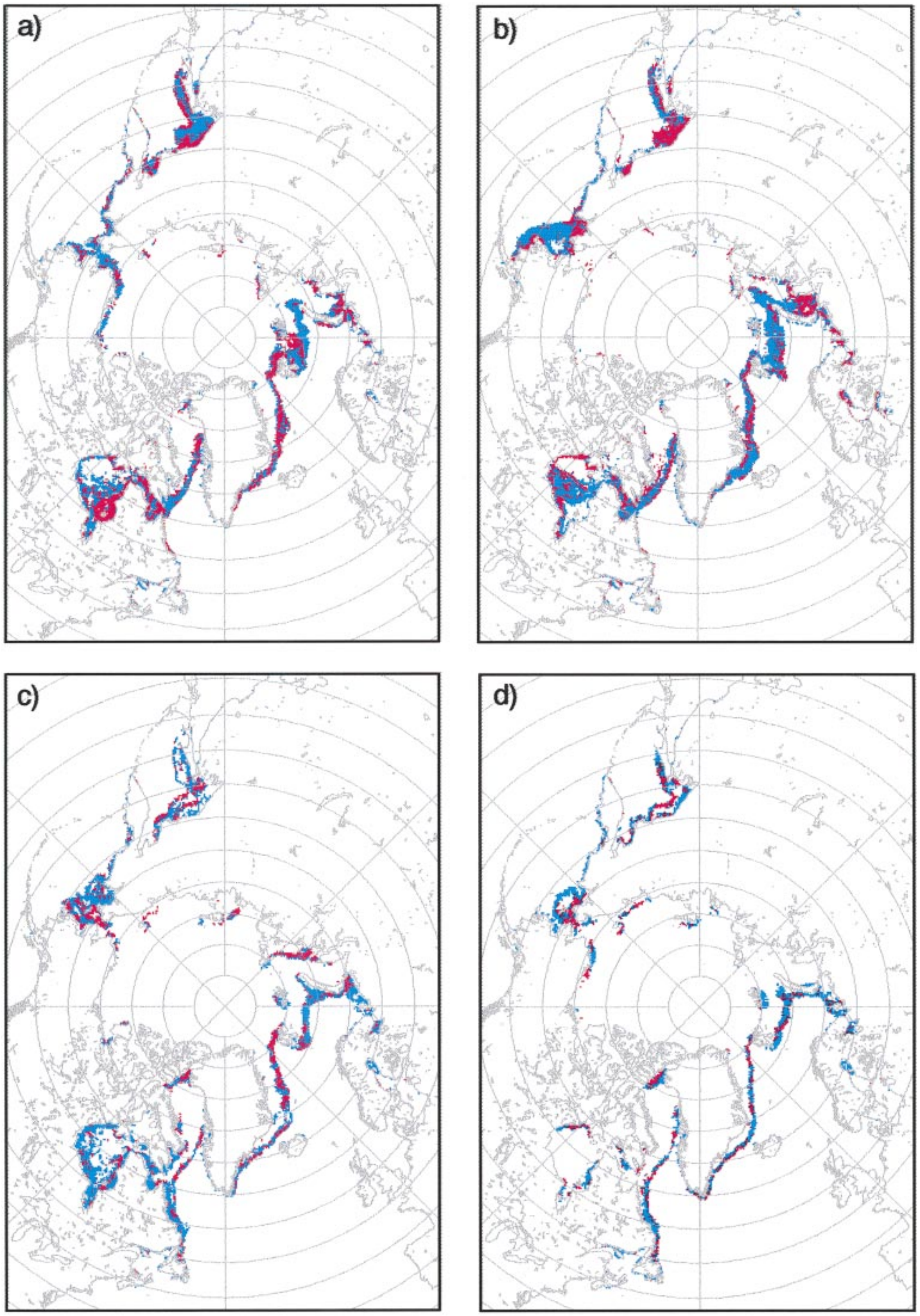

FIG. 7. Spatial distributions of SS_{cp}^{xy} for (a) Dec 2000, (b) Dec 2001, (c) May 2001, and (d) May 2002. Skillful forecasts ($SS_{cp}^{xy} > 0.1$) are denoted by blue and nonskillful forecasts ($SS_{cp}^{xy} < 0.1$) are denoted by red. Regions with unchanged sea ice or ocean conditions are denoted by white.

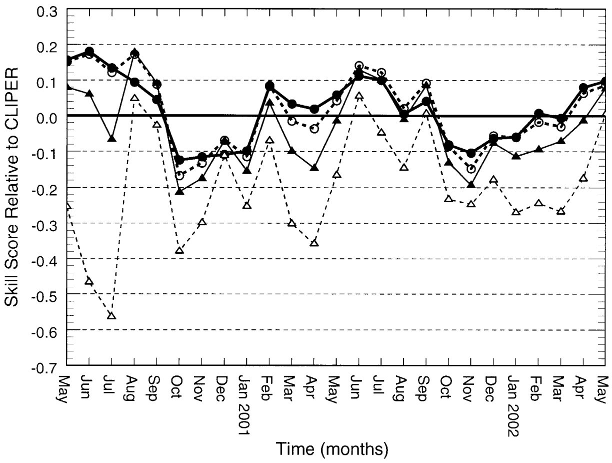

FIG. 8. Skill scores relative to CLIPER for the period May 2000–May 2002 based on a ±15% cutoff value. The bold solid and dashed lines with circles denote the 24- and 48-h skill scores, respectively. The solid and dashed lines with triangles denote the 72- and 120-h skill scores, respectively.

The skill scores relative to CLIPER (Fig. 8) based on a ±15% threshold are comparable to the skill scores based on a ±5% threshold (Fig. 3c) during spring and summer. However, the ±15% threshold skill scores show a marked decrease in skill at all time lags during the winter months. This suggests that PIPS has particular difficulties in predicting large changes in sea ice concentration during the freeze-up months.

b. Threat index

1) THREAT INDEX FOR A 5% DECREASE IN ICE CONCENTRATION

For the $T = 24$ h forecast, the number of correct forecasts of decreased ice concentration (hits; Fig. 9a) shows a strong annual progression in both years with the number of hits peaking during the summer months of May, June, and July. In contrast, the number or correct rejections reaches its maximum value during the winter months. The seasonal changes in these statistics merely reflect the way in which the problem was posed; namely, that a 5% decrease in sea ice concentration is termed a hit, which most frequently occurs during the spring/summer. Conversely, the correct rejection corresponds to a 5% or greater increase in sea ice concentration, which most frequently occurs during the winter. The total correct forecasts (hits + correct rejections; Fig. 10) vary between 20% and 30% and generally indicate better performance during the winter months. The breakdowns by category for the $T = 48$, 72, and 120 h forecasts are shown in Figs. 9b–d. These forecasts show the same general seasonal structure as the $T = 24$ h forecasts, but the number of correct forecasts decreases to ~13% at $T = 120$ h (Fig. 10). At each forecast time the misses and false alarms complete the contingency table and account for the remaining 70%–90% of the possible forecast outcomes. It is worth noting that the relative frequencies of hits and correct rejections for May 2000 are substantially less than the values reported by Van Woert et al. (2001). The reason for this discrepancy is that Van Woert et al. (2001) did not categorize the miss and false alarm forecasts falling within the strips denoted by $\pm c$ along the **O** and **F** axes in Fig. 2. These points are included in this study's statistics, resulting in a significant increase in the total number of points considered and a corresponding decrease in the fraction of hits and correct rejections observed.

The misses and false alarms show substantially less seasonal variability than the correct forecasts. However, at all forecast times and months, except during the fall, there is a slight tendency for more false alarms than misses. This gives rise to a slight bias in the forecast (Fig. 11) [forecasted frequency of decreasing ice cover (hits + false alarms) greater than the observed frequency of decreasing ice cover (hits + misses); see Fig. 2 for the definition of the quadrants]. That is, the model predicts more large decreases than actually occur except during the freeze-up months. Alternatively, during the fall months, there is a slight tendency for the forecasted frequency of increasing ice cover (correct rejections + misses) to be greater than the observed increase in ice cover (correct rejections + false alarms) (figure not shown). Thus, the model also tends to predict slightly more large increases in ice cover than actually occur during the freeze-up months.

2) THREAT INDEX FOR MELT-OUT CONDITIONS

The melt out (hits) for the $T = 24$ h forecast ranges between 2% and 8%, with the largest scores observed during the summer months and the lowest scores observed during the winter months (Fig. 12a). As was the case for the previous example of predicting 5% changes, this result is in accord with the anticipated seasonal freeze/melt cycle. Combined melt-out and freeze-up values (hits + correct rejections) vary between 8% and 10% (Fig. 12b) with misses and false alarms being roughly equally represented in the distribution (Fig. 12a). As such the freeze/melt forecasts appear to be relatively unbiased. Finally, similar to results of the other statistics, the forecast skill decreases with increasing time lag. In particular, for $T = 120$ h the skill barely exceeds 3%. It is concluded therefore that beyond 24 h PIPS has difficulty predicting concentration changes in the MIZ.

5. Summary and discussion

In this study we have assessed the forecast skill of the Polar Ice Prediction System sea ice concentration fields for the 25-month period May 2000–May 2002 using climatology, persistence, and combined climatology–persistence fields as the reference states. Measures of forecast skill were adapted from the meteorological

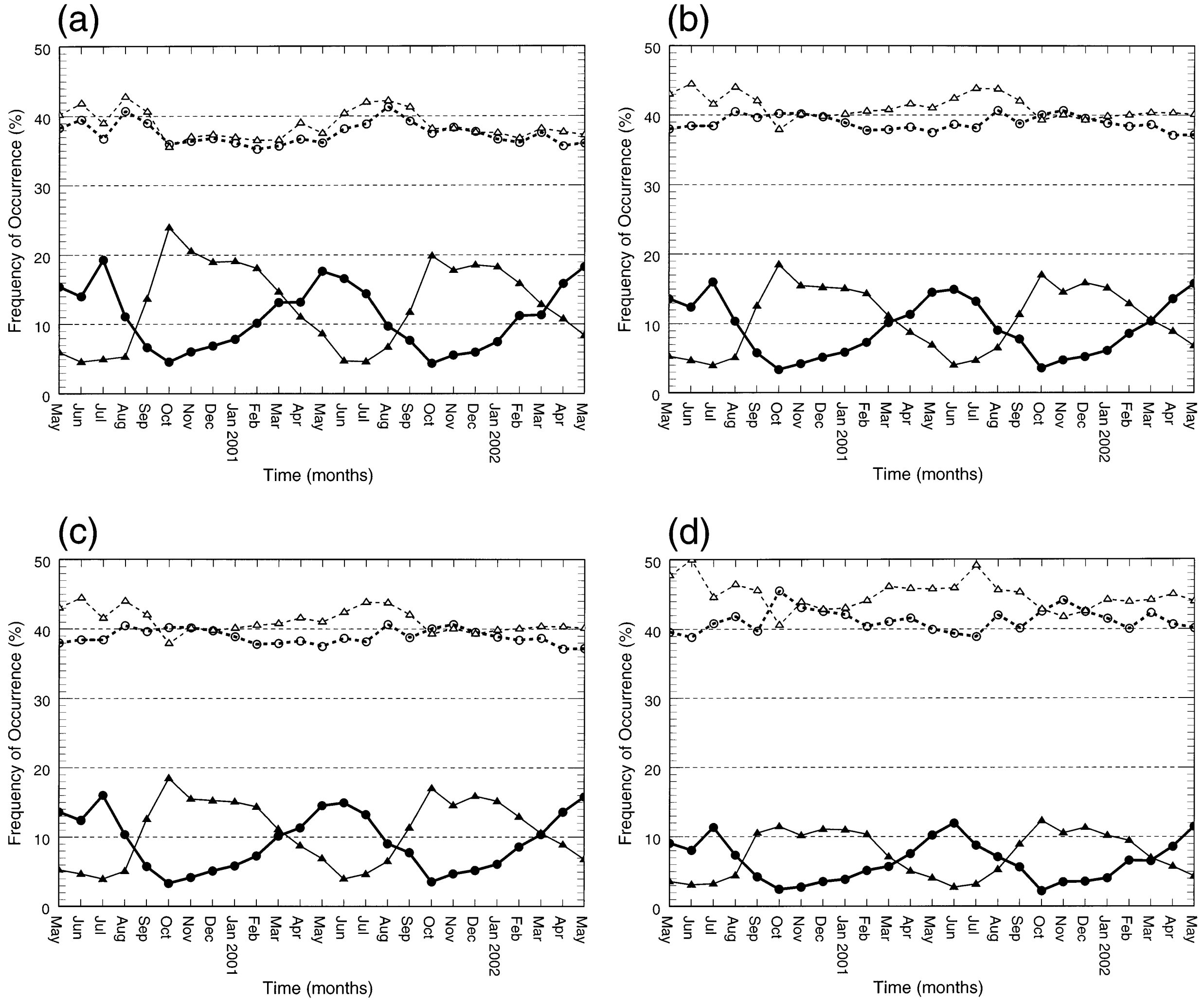

FIG. 9. Threat indexes for a forecasted >5% decrease in sea ice concentration at (a) 24-, (b) 48-, (c) 72-, and (d) 120-h lags. Hits are denoted by a bold line with solid circles, correct rejections by a line with solid triangles, misses by a bold dashed line with open circles, and false alarms by a dashed line with open triangles.

literature with the mean-square error being used as the measure of accuracy. Scores in the range 0.0–1.0 indicate the positive impact of the model relative to the reference fields. When referenced to climatology, the $T = 24$ h skill score was ~0.85 indicating that PIPS provides a significantly better forecast than climatology during all months examined. The high skill score is most likely due to the daily reinitialization of PIPS with SSM/I data, which generally gives a more accurate representation of current sea ice conditions than is obtainable from climatology.

Persistence and combined climatology–persistence are two other common reference states for forecast evaluation. PIPS exceeds a persistence forecast during the non-freeze-up months of March–October with peak $T = 24$ h skill scores of ~0.2, corresponding to a correlation coefficient of ~0.45. In contrast, using a reference field comprising ~95% persistence and ~5% climatology, peak $T = 24$ h forecasts are ~0.1. This reduction in skill score referenced to a combination of persistence and climatology is consistent with meteorological studies, which have shown that a linear combination of the persistence and climatological fields provides the best possible reference state for model evaluation. One curious finding of this analysis is that the 120-h forecast skill is actually better than the 24-h forecast relative to a persistence forecast during many months of the year. One possible explanation for this result is that the model is attempting, over time, to correct deficiencies in the initialization field, which is based on the direct insertion of SSM/I-derived ice concentration data into the model. This suggests that additional work is needed to develop dynamically sound initialization fields for ice forecast models. Finally, this analysis suggests that during the winter months when PIPS performs poorly, a linear combination of climatological

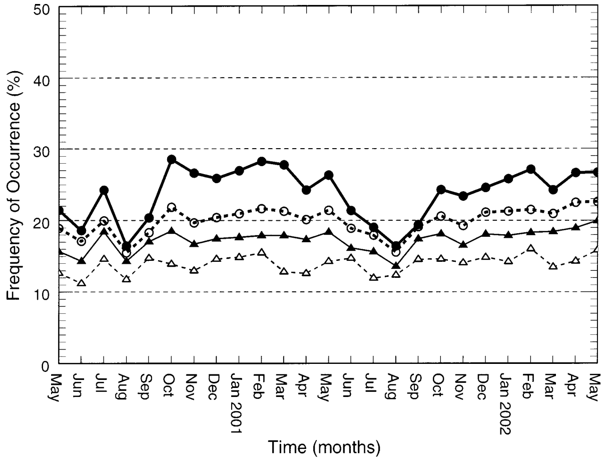

FIG. 10. Total correct (hits + correct rejections) forecasts at 24 (bold line with solid circles), 48 (bold dashed line with open circles), 72 (line with solid triangles), and 120 h (dashed line with open triangles).

and the current ice conditions provides the current best possible short-term forecast.

The previously discussed parametric statistics give a rigorous assessment of PIPS's ability to forecast specific ice concentration values. At times, however, simply knowing whether ice conditions are likely to change significantly at some future date could provide useful planning information, independent of whether the model correctly forecasts the specific concentration value. For example, a forecast of decreasing ice concentration in the vicinity of a ship beset in ice might indicate that it could be released from the ice. In contrast, if the ice concentration is predicted to increase, one might logically begin to plan some form of a rescue operation. A ''threat index'' is ideally suited to assessing the skill of binary yes–no questions such as these. Using this criterion for forecast success, it was found that PIPS correctly forecasts sea ice conditions ~25% of the time, with the best forecasts occurring during the winter months.

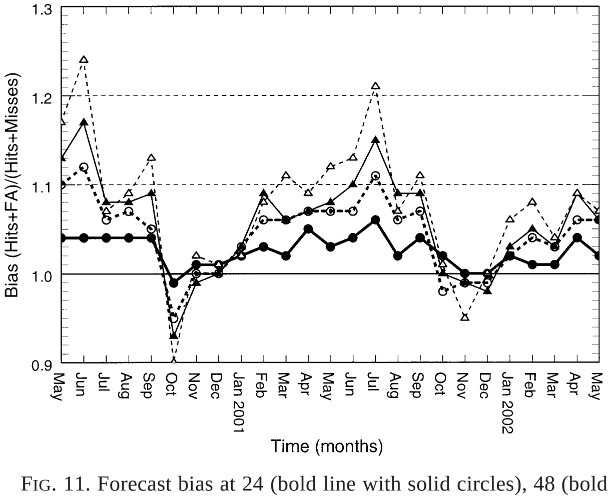

FIG. 11. Forecast bias at 24 (bold line with solid circles), 48 (bold dashed line with open circles), 72 (line with solid triangles), and 120 h (dashed line with open triangles).

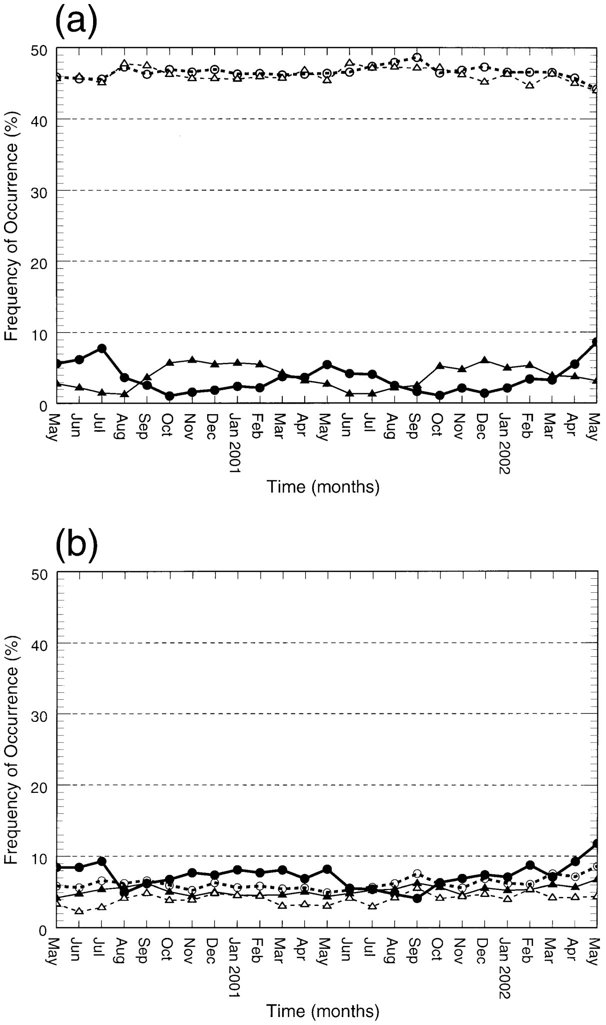

FIG. 12. (a) Threat indexes for forecasted melt of sea ice at 24-h lag. Hits are denoted by a bold line and solid circles, correct rejections by a line and solid triangles, misses by a bold dashed line with open circles, and false alarms by a thin line with open triangles. (b) Total correct forecasts at 24 (bold line with solid circles), 48 (bold dashed line with open circles), 72 (line with solid triangles), and 120 h (dashed line with open triangles).

The threat index methodology is also well suited to assessing the ability of PIPS to forecast melt and freeze conditions along the ice edge. Not surprisingly, most of the sea ice variability within PIPS is located at the MIZ. In this study it was found that PIPS correctly forecasted melt–freeze changes at the MIZ <10% of the time for all forecasts studied. Preller and Posey (1996) noted regional discrepancies in PIPS that they attributed to deficiencies in the parameterization of shelf processes, the treatment of sea ice as a continuum rather than separate floes, and inaccuracies in the atmospheric forcing. Van Woert et al. (2001) support these general conclusions, suggesting that deficiencies in the atmospheric model driving PIPS and/or uncertainties in the Arctic Ocean model that forms the basis of PIPS are significant sources of error. This study suggests further that PIPS probably does not properly parameterize processes occurring at the MIZ.

This study has focused on the evaluation of the PIPS. However, the skill score formulations discussed here are broadly applicable to a wide range of ocean and ice forecast models. As the international community moves forward to establish ‘‘operational oceanography’’ programs, increased effort will need to be placed on evaluating these systems. This study highlights some of the difficulties that one might expect to encounter in trying to develop a forecast system that has positive impact when the geophysical fields are highly persistent.

Acknowledgments. Ted Maksym and two anonymous reviewers provided valuable suggestions for improving the original draft of this manuscript. The SSM/I data from 1 June 1995 through 31 March 2001 were provided by the EOSDIS NSIDC Distributed Active Archive Center (NSIDC DAAC), University of Colorado. The remainder of the SSM/I data were provided by NOAA/NESDIS. Mike Chase kindly provided the sea ice climatology fields used in this study. We thank Dennis Conlon and Tom Curtin of the Office of Naval Research for their support of this work. This paper is approved for public release, distribution unlimited.

REFERENCES

Arpe, K., A. Hollingsworth, M. S. Traction, A. C. Lorenc, S. Uppala, and P. Kallberg, 1985: The response of numerical weather prediction systems to FGGE level IIb data. Part II: Forecast verifications and implications for predictability. *Quart. J. Roy. Meteor. Soc.,* **111,** 67–101.

Atlas, R., M. Ghil, and M. Halem, 1981: Reply. *Mon. Wea. Rev.,* **109,** 201–204.

Beesley, J. A., C. S. Bretherton, C. Jakob, E. L Andreas, J. M. Intieri, and T. A. Uttal, 2001: A comparison of cloud and boundary layer variables in the ECMWF forecast model with observations at the Sheba Ice Camp. *J. Geophys. Res.,* **105,** 12 337–12 349.

Brier, G. W., and R. A. Allen, 1951: Verification of weather forecasts. *Compendium of Meteorology,* T. Malone, Ed., Amer. Meteor. Soc., 841–848.

Cheng, A., and R. H. Preller, 1996: The development of an ice–ocean coupled model for the Northern Hemisphere. NRL Rep. NRL/FR/7322—95-9627, Naval Research Laboratory, Stennis Space Center, MS, 65 pp.

Cox, M., 1984: A primitive equation, 3-dimensional model of the ocean. GFDL Ocean Group Tech. Rep. 1, Geophysical Fluid Dynamics Laboratory, Princeton, NJ, 161 pp.

Efron, B., 1979a: Bootstrap methods: Another look at the jackknife. *Ann. Stat.,* **7,** 1–26.

——, 1979b: Computers and the theory of statistics: Thinking the unthinkable. *SIAM Rev.,* **21,** 460–480.

Flato, G. M., and W. D. Hibler III, 1992: Modeling pack ice as a cavitating fluid, *J. Phys. Oceanogr.,* **22,** 626–651.

Ghil, M., 1979: Time-continuous assimilation of remote sensing data and its effect on weather forecasting. *Mon. Wea. Rev.,* **107,** 140–171.

Grumbine, R. W., 1998: Virtual floe ice drift forecast model intercomparison. *Wea. Forecasting,* **13,** 886–890.

Heirtzler, J. R., Ed., 1985: Relief of the surface of the Earth. Rep. MGG-2, National Geophysical Data Center, Boulder, CO.

Hibler, W. D., III, 1979: A dynamic–thermodynamic sea ice model. *J. Phys. Oceanogr.,* **9,** 815–846.

——, 1980: Modeling a variable thickness sea ice cover. *Mon. Wea. Rev.,* **108,** 1943–1973.

Hogan, T. F., and T. E. Rosmond, 1991: The description of the Navy Operational Global Atmospheric Prediction System's spectral forecast model. *Mon. Wea. Rev.,* **119,** 1786–1815.

Hollinger, J. P., 1991: DMSP special sensor microwave/imager calibration/validation—Final report. Vol. 2, Naval Research Laboratory, Washington, DC, 310 pp.

Hollingsworth, A., K. Arpe, M. Tiedtke, M. Capaldo, and H. Savijarvi, 1980: The performance of a medium-range forecast model in winter—Impact of physical parameterizations. *Mon. Wea. Rev.,* **108,** 1736–1773.

Koblinsky, C. J., and N. R. Smith, Eds., 2001: Observing the oceans in the 21st century. *Proceedings of the First International Conference on Ocean Observations for Climate (OCEANOBS99),* GODAE Project Office, Melbourne, Australia.

Levitus, S., 1982: *Climatological atlas of the world ocean.* NOAA Prof. Paper 13, 173 pp. and 17 microfiche.

Mace, G. G., C. Jakob, and K. P. Moran, 1998: Validation of hydrometeor occurence predicted by the ECMWF model using millimeter wave radar data. *Geophys. Res. Lett.,* **25,** 1645–1652.

Meier, W. N., M. L. Van Woert, and C. Bertoia, 2001: Evaluation of operational SSM/I concentration algorithms. *Ann. Glaciol.,* **33,** 102–108.

Murphy, A. H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.,* **116,** 2417–2424.

——, 1992: Climatology, persistence, and their linear combination as standards of reference in skill scores. *Wea. Forecasting,* **7,** 692–698.

——, and H. Daan, 1985: Forecast evaluation. *Probability, Statistics and Decision Making in the Atmospheric Sciences,* A. H. Murphy and R. W. Katz, Eds., Westview Press, 379–437.

——, and E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.,* **117,** 572–581.

Partington, K. C., 2000: A data fusion algorithm for mapping sea-ice concentrations from Special Sensor Microwave/Imager data. *IEEE Trans. Geosci. Remote Sens.,* **38,** 1947–1958.

Pinardi, N., and J. Woods, Eds., 2002: *Ocean Forecasting: Conceptual Basis and Applications.* Springer, 472 pp.

Posey, P. G., and R. H. Preller, 1994: Operational use of SSM/I ice concentration in the initialization of a coupled ice–ocean model. *Proc. IGARSS '94*, Pasadena, CA, California Institute of Technology, 1231–1233.

Preller, R. H., 1992: Sea ice prediction: The development of a suite of sea-ice forecasting systems for the Northern Hemisphere. *Oceanography*, **5,** 64–68.

——, 1999: Prediction in ice-covered shallow seas. *Coastal Ocean Prediction,* Vol. 56, *Coastal and Estuarine Studies,* Amer. Geophys. Union, 405–441.

——, and P. G. Posey, 1989: The Polar Ice Prediction System—A sea ice forecasting system. NORDA Rep. 212, Naval Research Laboratory, Stennis Space Center, MS, 42 pp.

——, and ——, 1996: Validation test report for a Navy sea ice forecasting system: The Polar Ice Predication System 2.0. NRL Rep. NRL/FR/7322—95-9634, Naval Research Laboratory, Stennis Space Center, MS, 31 pp.

——, J. E. Walsh, and J. A. Maslanik, 1992: The use of satellite observations in ice covered simulations. *Microwave Remote Sensing of Sea Ice, Geophys. Monogr.,* No. 68, Amer. Geophys. Union, 385–403.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in FORTRAN 77.* 2d ed. Cambridge University Press, 933 pp.

Pritchard, R., S. Mueller, A. C. Hanzlick, and Y.-S. Yang, 1990: Forecasting Bering Sea ice edge behavior. *J. Geophys. Res.,* **95,** 775–788.

Steffen, K., J. Key, D. Cavalieri, J. Comiso, P. Gloersen, K. St. Germain, and I. Rubinstein, 1992: The estimation of geophysical parameters using passive microwave algorithms. *Microwave Remote Sensing of Sea Ice, Geophys. Monogr.,* No. 68, Amer. Geophys. Union, 201–231.

Stephenson, D. B., 2000: Use of the ''odds ratio'' for diagnosing forecast skill. *Wea. Forecasting,* **15,** 221–232.

Van Woert, M. L., W. N. Meier, C.-Z. Zou, J. A. Beesley, and P. D. Hovey, 2001: Satellite validation of the May 2000 sea ice concentration fields from the Polar Ice Prediction System. *Can. J. Remote Sens.,* **27,** 443–456.

Walsh, J. E., W. D. Hibler, and D. Ross, 1985: Numerical simulation of Northern Hemisphere sea ice variability, 1951–1980. *J. Geophys. Res.,* **90,** 4847–4865.

Weaver, R., C. Morris, and R. G. Barry, 1987: Passive microwave data for snow and ice research: Planned products from the DMSP SSM/I system. *Eos, Trans. Amer. Geophys. Union,* **68,** 769, 776–777.