Research Department

Technical Memorandum No. 307

Quantitative Precipitation Forecasts over the United States by the ECMWF Ensemble Prediction System

by

Steven L. Mullen[1]

Roberto Buizza[2]

[1] Institute of Atmospheric Physics, University of Arizona, Tucson, Arizona

[2] European Centre for Medium-Range Forecast, Reading, RG2 9AX, United Kingdom

Accepted for publication by Monthly Weather Review

June 2000

ABSTRACT

The performance of the ECMWF Ensemble Prediction System (EPS) is assessed for probabilistic forecasts of 24-h accumulated precipitation over the eastern United States. Daily forecasts for the period 1 January 1997 to 31 January 1999 are verified for projections of one to 10 days. Verification is performed separately for the cool and warm seasons, and the impact of changes to the EPS that occurred during the study period is assessed. Analyses of rain gauge data from the River Forecast Centers of NOAA are used for verification. Skill is measured relative to long-term climatic frequencies, and the statistical significance of differences in the accuracy and skill among forecasts is estimated.

Overall, EPS forecasts are more skillful during the winter than the summer. The EPS produces significantly skillful forecasts to past one week for a threshold of 1 mm in both seasons. Accuracy decreases as the threshold increases, until forecasts of 50 mm are not significantly skillful at 1 day. The implementation of evolved singular vectors in the EPS appears to have minimal impact on skill during the summer. The addition of both evolved singular vectors and stochastic processes in the EPS appears to improve short-range performance for thresholds between 1-20 mm during the winter, but results for higher thresholds (50 mm) are equivocal.

1. INTRODUCTION

Improving the quality of quantitative precipitation forecasts (QPFs) is a primary goal of operational prediction centers and a major challenge facing the research community (e.g. *Fritsch* 1998). The skill of QPFs has lagged behind that for other weather parameters (e.g. *Sanders* 1986), and precipitation forecasts continue to deteriorate more rapidly than forecasts of any other sensible weather element. Thus, it is extremely desirable to examine the degree to which relatively new forecast paradigms, such as ensemble prediction, can improve the skill of QPFs and probabilistic quantitative precipitation forecasts (PQPFs).

Ensemble predictions have been run at two operational prediction centers, the European Centre for Medium-Range Weather Forecasts (*ECMWF, Palmer et al.* 1993) and the National Center for Environmental Prediction (*NCEP, Tracton and Kalnay* 1993), since December 1992, and shortly thereafter at other operational centers (e.g. *Pauley et al*. 1995, *Houtekamer et al.* 1996). Operational implementations of mesoscale, short-range ensemble systems are planned in the near future (*Brooks et al.* 1995, *Tracton and Du* 1998). Verifications of probabilistic predictions for various forecast parameters have established the skill and utility of operational ensemble systems for planetary and synoptic scale features (*Molteni et al.* 1996, *Buizza* 1997ab, *Toth et al.* 1997, 1998, *Stensrud et al.* 1999). Verification of PQPFs for medium-range ensembles (*Buizza et al.* 1999a, *Eckel and Walters* 1998) and experimental short-range ensembles (*Du et al.* 1997, *Hamill and Colucci* 1998, *Stensrud et al.* 1999) has demonstrated that ensemble predictions can be more beneficial than single deterministic forecasts at higher resolutions. In regards to QPF, *Buizza et al.* (1999a) found for the ECMWF system that skill in ensemble predictions over Europe persisted into the medium range for low thresholds, but not for high ones. Skill was lower in summer than in winter, although it was somewhat improved by a recent increase in the ensemble size and model resolution.

The focus of this study is the statistical verification of the performance of the ECMWF Ensemble Prediction System (EPS) for PQPF. This study extends the earlier verification by *Buizza et al.* (1999a) for Europe in several respects. First, two major system upgrades have been implemented in the EPS since their work. They verified 12 h accumulations on a 2.5° x 2.5° grid using 12-24 h prognoses from the ECMWF deterministic high resolution forecast as a proxy of the observed field. Here, model verification of 24 h accumulations is performed against analyses based on reports of 24 h rainfall. Verification is performed over the contiguous United States on a finer 1.25° x 1.25° grid, a spacing very close to the reduced linear grid Gaussian grid of the model (1.125°). Because heavy precipitation events typically pose the severest threat to public safety and commerce (*Fritsch et al.* 1998), EPS performance for a higher threshold (50 mm per day) is also examined in this paper.

In the first part of this paper, the aggregate EPS performance for PQPF for the cool and warm seasons is contrasted. Next comparisons of 1997 and 1998 summers and of the 1997-98 and 1998-99 winters are performed to assess the impact of system changes on PQPF. Finally, case studies from summer 1998 and winter 1998-99 are presented to identify common error characteristics for heavy precipitation events.

This paper is organized as follows. Section 2 describes the configuration of the EPS, while section 3 summarizes verification procedures. Section 4 gives aggregate statistics for the cool and warm seasons, and Section 5 compares performance between the 1997 and 1998 warm seasons and between 1997-98 and 1998-99 cool seasons. Section 6 describes EPS performance for selected cases with observed heavy (>50 mm) precipitation. Section 7 presents an analysis of our results in light of other works, while section 8 highlights the major findings of the study.

2. Overview of the ECMWF Ensemble Prediction System

The ECMWF EPS contains 51 members, an unperturbed control forecast and 25 pairs of twin forecasts with positive and negative initial perturbations. The resolution of the EPS (*Buizza et al.* 1998) is spectral with triangular truncation at total wave number 159 and 31 layers (TL159L31, where the subscript “L” stands for the “linear grid” option, whereby the computational grid is the same as the standard “quadratic” grid used for the T106 model version). Initial perturbations are constructed from linear combinations of the leading singular vectors for a total energy norm and optimization time of 48 h (*Buizza et al.* 1998). The singular vectors are computed at low resolution (T42L31) with a linear and adjoint version of the model, and are defined only in terms of temperature, vorticity, divergence and surface pressure. The only physical processes included in the singular vector computation are simple surface drag and vertical diffusion; moist processes are not included.

The ECMWF EPS experienced two major system changes during the period 1 January 1997 to 31 January 1999 (*Buizza et al.* 1999b). On 26 March 1998, singular vectors that also sample growing instabilities during the data assimilation period (*Barkmeijer et al.* 1999) were implemented. On 21 October 1998, the simulation of random model errors related to the sub-grid scale uncertainty of physical parameterizations (*Buizza et al.* 1999c) was introduced. The impacts of these systems upgrades on EPS performance for PQPF are also assessed in this paper.

3. Verification procedures

3.1 Verification data and analyses

The data used for verification are 24 h accumulations of precipitation collected by the River Forecast Centers (RFC) of the National Weather Service. The RFC network includes approximately 5000 stations that report 24 h accumulations valid at 1200 UTC. The RFC data are used by NCEP to verify their precipitation forecasts. Our procedure for mapping the station data onto the 1.25° x 1.25° grid also follows NCEP procedures.

The so-called "box averaging" technique is used to grid the RFC precipitation data. Observations are assigned to the nearest grid box, and gridded value is the arithmetic mean of all observations within each box (*Messinger* 1996, p.2638). Figure 1 shows the spatial distribution of station density for the RFC data, averaged over the entire study period. Large spatial variations in average density are evident. A large swath of high densities (10 or more stations per grid box) covers most of U.S. from Central Texas northeastward into Maine. Moreover, the general pattern of highest station density to east of the Rocky Mountains and lowest density over the Great Basin and the western U.S. does not change greatly from day to day. The consistency of the pattern indicates a consistent daily bias in the spatial distribution of stations across the continent. It is reasonable to assume that, all other factors being equal, the accuracy of the verifying analyses should increase with station density. However, empirical evidence on the relationship amongst grid spacing, station density and distribution within a grid box, and analysis fidelity is lacking (e.g. *Fritsch et al.* 1998). As a reasonable compromise between analysis accuracy and the number of model grid boxes to retain, verification is performed only at grid points with 5 or more stations.

The RFC station reports that are archived by NCEP are not subjected to quality control (QC). Subjective synoptic evaluation of a few analyses (~90 case days) of RFC data revealed a small number of suspected spurious reports. For that reason, a simple QC procedure was performed before mapping the RFC data onto the uniform grid. Any report greater than 25.4 mm (1.0") was subjected to a QC-check. If such a report exceeded the average precipitation for all stations within a 100-km radius of the said station by more than 10 standard deviations, it was excluded from the box averaging analysis. This QC procedure typically eliminated an average of 1-2 stations per day, and at most 3-4 stations per day. Admittedly, the QC procedure might occasionally remove some accurate reports of heavy precipitation, especially during the warm season when convection becomes prevalent and synoptic-scale forcing weakens. Subjective evaluation of analysis with and without QC indicated that isolated reports of high precipitation were often eliminated during cool season, which could reflect such errors as snowfall or snow depth being erroneously recorded as water equivalent.

Two precipitation analyses for 1200 UTC 25 January 1997 are show in Fig. 2 that illustrate typical behavior of the 5-station minimum and the QC procedure. The main effect of the 5-station stipulation is to reduce the area of the verification region; for example, note the absence of shading over the Central and Upper Plains States in Fig. 2b. The QC procedure mitigates the impact of an isolated report of heavy precipitation (16.14 inches or ~64 mm). This adjustment can be seen over western Illinois where a greater than 50 mm value for a grid point (Fig. 2a) becomes ~10 mm (Fig. 2b) once the questionable report is eliminated from the analysis. Examination of 6 hourly surface weather maps and individual precipitation reports clearly indicate that a 50 mm amount is utterly wrong. On the other hand, regions of heavy precipitation with two or more corroborating reports within the close proximity of each other are deemed valid. Note the agreement between the two panels for the region of heavy rain over the Gulf Coast States.

Archives of 24 h precipitation (National Weather Service Summary of the Day NCDC TD 3210) from the National Climate Data Center (NCDC) were used to compute climatological frequencies of precipitation for the four thresholds. Station reports, accurate to 0.01" (0.254 mm), were obtained for 353 sites in the contiguous United States. The sites denote stations for which NCEP produces forecast guidance with Model Output Statistics (e.g. *Carter et al.* 1989). The period of record varies from station to station, but it exceeds 25 years for all stations and 40 years for the majority of stations. For each station, the frequency of the occurrence of accumulations of at least 1, 10, 20 and 50 mm were computed. The box averaging technique was not used to grid the climatological frequencies since a large number of the boxes would not contain a station, especially in the western United States. Fields of precipitation frequency for each threshold were instead mapped on the uniform grid by applying a two-pass version of the *Barnes* (1973) objective analysis scheme with half-power cutoff of ~600 km. For each verification time, the climatological frequency was defined as a 31-day mean centered on the verification day.

3.2 Accuracy Measures

A single number is usually inadequate for evaluating all the desired information about the performance of an analysis/forecast system (*Murphy and Winkler* 1987, *Murphy* 1991). Each verification measure provides unique information on performance. For that reason, the utility of EPS forecasts will be evaluated by several measures. Following *Buizza et al.* (1999a), EPS performance is primarily evaluated thorough application of reliability diagrams, *Brier* (1950) scores and signal detection theory (*Mason* 1982, *Stanski et al.* 1989). The Ranked Probability Score (*Epstein* 1969, *Murphy* 1971) and verification of rank (*Anderson* 1996, *Hamill and Colucci* 1998) are also employed. The reader is referred to overviews by *Stanski et al.* (1989) and *Wilks* (1995), and references therein, for thorough descriptions of the verification methods. Brief descriptions of accuracy measures and definitions of acronyms for them are given in the Appendix.

3.3 Statistical significance testing

The statistical problem is to determine whether accuracy for the EPS forecasts is significantly better than a baseline forecast (e.g. climatology), or whether changes in EPS configuration from one year to the next, or between the EPS forecasts and idealized perfect model configurations, produce a significant change in the accuracy or skill measures. The significance of differences between RMSE's, Brier Skill Scores, areas under Relative Operating Characteristic (ROC) curves and rank frequencies is judged by non-parametric resampling. The use of bootstrap techniques to estimate statistical significance has clear advantages over parametric tests such as the student's t-test (e.g. *Livezey and Chen* 1983; *Wilks* 1995, pg. 145-150, and references within). Some advantages of resampling techniques are that i) an *a priori* assumption of the background distribution of the sample populations (e.g. equal variances) is not required, ii) an *a priori* estimate of the number of degrees of freedom is not required, and iii) differences between any metric can be tested. The resampling technique that is used in this study to estimate the statistical significance of differences and confidence bounds is also described in the end of the Appendix.

4. Overall performance during the cool and warm seasons

4.1 Spatially averaged verification statistics

The results described in this section are based on every 1200 UTC forecast cycle during the period 1 January 1997 to 31 January 1999, inclusive. Verification was performed for the U.S. region 102.5° W and eastward, where the RFC station density tends to run highest. To facilitate comparison with the verification of the NCEP ensemble system by *Eckel and Walters* (1998), the cool/winter season is defined as the five-month period of November through March, while the warm/summer season is defined as the five-month period of May through September. Transition months of April and October are not included in the results.

Brier skill scores (BSS) for all precipitation thresholds are shown in Fig. 3 for winter and summer, along with an estimate of 90% confidence bounds (Appendix). In winter, EPS forecasts for amounts 10 mm or less remain significantly skillful relative to long-term climatology for more than one week, while forecasts for 20 mm show significant skill at days 1 and 3. The 50 mm category has no skill even at 1 day. Not surprisingly, summer forecasts possess less skill. Only 1 mm retains skill longer than one week. The 10 mm threshold loses skill after just one day, while both 20 mm and 50 mm are even unskillful at day 1.

A curious feature of the BSS curves is the minimum in skill at day 2 for the 20 and 50 mm thresholds, especially during summer when BSS values dip to -0.2 and -0.4 respectively. Day 2 is the optimization time for the singular vectors that form the basis functions for the initial perturbation perturbations for the EPS (*Buizza et al.* 1998). This, and the fact that the EPS initial perturbations are computed without moist processes, suggests that the minimum might be related in part to the perturbations not being optimized for short-range forecasts of precipitation. The T42L31 perturbations also lack the resolution to sample scales typical of convective systems, and do not include an explicit constraint on precipitation or variables closely related to convection such as moisture, vertical velocity and vertical stability (*Doswell* 1987). Thus, it is not surprising that the EPS perturbations, which are not explicitly designed to maximize dispersion of precipitation, would not make precipitation forecasts skillful for an ensemble system that lacks dispersion (as shown later). Moreover, a “spin-up” problem in the convection scheme that manifests itself most strongly with heavy precipitation events might also be a contributing factor to the early minimum in the BSS.

Further insight into the behavior of the EPS can be obtained from reliability diagrams (Figs. 4 and 5) and decomposition of the Brier Score (Figs. 6 and 7). A ubiquitous tendency of the EPS, for all thresholds during both seasons, is over forecasting of the likelihood of precipitation. Figures 4 and 5 indicate that forecast probabilities are systematically larger than conditional probabilities of occurrence. While the reliability tends to become poorer as the threshold increases, there is strong tendency for the reliability curves to increase monotonically. The primary exception is the jagged curves for 50 mm, a clear sign of an insufficient sample size (e.g. *Wilks* 1995, pg 266-267). Thus, higher forecast probabilities do tend to correspond with higher observed frequencies for thresholds for which large sample sizes are available.

It is interesting to note that the reliability curves for even the most skillful forecasts hover near the no-skill line. This indicates that EPS performance would benefit greatly from even slight improvements in either reliability or resolution. Decomposition of the Brier Score (Figs. 6 and 7) shows that the resolution term runs half the size of the uncertainty term (sample climatology) only for light thresholds during winter. The short-range (1 to 3 day) forecasts for the light amounts are skillful because they possess the desirable attributes of resolution and sharpness (*Wilks* 1995, pg 237): the vast majority of short-range 1 mm forecasts are properly placed into either the driest or the wettest category. Both the resolution and sharpness decrease as the threshold increases and the forecast projection lengthens, as would be expected. On the other hand, the reliability term tends to decrease (acts to improve the BSS) with projection, and reaches its highest value (worse) at day 2 for 20 mm and 50 mm during both summer and winter.

Figure 8 shows the evolution of area under the ROC curve for all precipitation categories for the winter and summer seasons. The wintertime values indicate a useful ability to discriminate precipitation events to day 10. Of particular interest is the ROC area for the heaviest category of 50 mm. The 90% confidence bounds for wintertime curve remain above the 0.5 value to day 10. This indicates that the EPS can readily discriminate heavy precipitation events during the winter, despite the lack of skillful Brier scores. The 90% percent confidence bounds for the summertime ROC values are also above 0.50 for all thresholds except 50 mm at day 10. ROC areas during summer run significantly below their wintertime counterparts for most projections, however, especially for the highest thresholds, a testament to the difficulty of forecasting heavy precipitation events during the warm season when synoptic forcing becomes weaker and convection becomes more dominant than during the cool season.

4.2 Spatial distribution of ranked probability skill score

When interpreting the aggregate results for the EPS, it is important to keep in mind that they represent area-averaged scores. Large spatial variations in skill and predictability exist in the EPS forecasts over the eastern United States. As a way to illustrate this spatial variability, distributions of the RPSS, averaged at individual grid points for the cool season and warm season, are shown Figs. 9 and 10, respectively. The RPSS values are based on all verification dates with grid points that have 5 or more RFC reports. Thus, grid boxes that contain less than 5 stations in Fig. 1 can show RPSS values in Figs 9 and 10.

The winter pattern (Fig. 9) shows the highest skill over the Pacific Northwest, windward of the Cascade Range. Skill scores remain above 0.5 along the coastline of Oregon and Washington out to day 7. Moreover, elevated skill along the Pacific Northwest is evident in distributions for all three winters (results not shown). Enhanced skill for a region immediately downwind of the North Pacific Ocean, especially in the short-range (1 to 3 days), might seem somewhat surprising because analysis uncertainty is greater over ocean than the continent. The result suggests that deficiencies in perturbation design and model formulation might be playing a more important role suppressing skill elsewhere than over the Pacific Northwest. Alternatively, wintertime precipitation might be inherently more predictable over the Pacific Northwest for the scales resolved by the EPS than elsewhere across the U.S. Wintertime skill is particularly low over the intermountain West with vast areas of no skill even at day 1, and to a lesser degree south of Lake Erie and Lake Ontario and over the upper Appalachian Mountains. Low skill over the intermountain West may reflect, in good part, greater uncertainty and errors in the verifying analyses since data density is so low in the region (Fig. 1). It is undoubtedly less of a factor downstream of the Great Lakes and over the Appalachians where data density is highest (Fig. 1). The reduction in skill could also reflect a model error, one that might be greatly ameliorated by an increase in horizontal resolution and a more accurate representation of surface heterogeneity.

The summer distribution (Fig. 10) indicates a tendency for reduced skill at low latitudes and enhanced skill in high latitudes. Skill over Texas and the southern Rocky Mountains is particularly poor, with loss of skill by day 1 over vast areas. The region of poor skill over central Texas contains decent data coverage (typically 10-20 stations per grid box), so analysis uncertainty and representativeness are probably less of a factor than over southern Rocky Mountains. A comparison of the RPSS pattern with distributions of thunderstorm frequency (*Wallace* 1975) and ground lightning flash density (*Huffines and Orville* 1999) suggests a negative correlation between two fields, another reflection of the difficulty of forecasting precipitation associated with convection.

5. Impact of system changes

5.1 Evolved singular vectors during the warm season

On 26 March 1998, evolved singular vectors that also sample growing errors during the data assimilation cycle were implemented in the EPS. As *Barkmeijer et al.* (1999) discuss, simultaneous sampling of past and future directions of error growth can improve the dispersion characteristics of the ensemble during the early periods of the forecast for fields such as height anomalies and temperature. The impact of this change on warm season precipitation can be assessed by comparing summers 1997 and 1998, where the summer season is defined as before.

Forecast evolutions of the BSS and ROC area are shown in Figs. 11 and 12, respectively, for the two summers. Overall, the inclusion of evolved singular vectors in 1998 had a minimal impact on EPS performance for precipitation forecasts; BSS values and ROC areas for the three smallest thresholds differ little between the two summers. Results (Table 1) of hypothesis testing, based on non-parametric permutation methods (Appendix), indicates that differences between BSS values are significant at the 5% level for the 1, 20, and 50 mm thresholds at day 1 (to day 2 for 20 mm), while differences between ROC areas are insignificant to day 6 for all thresholds. (We believe that the significant ROC values at day 8-9 for the 20 mm and day 7 for 50 mm likely reflect a well-known multiplicity problem discussed by *Wilks* (1995, p. 151-152).) Perhaps the most encouraging change is an apparent improvement in the 50 mm statistics for the summer with evolved singular vectors. The BSS at day 1 for summer 1998 is 0.2. Even the deleterious dip in the BSS at day 2 is noticeably less during 1998, owing to a major reduction in the BS_{rel} term from the prior year (Fig. 13 and 14).

Rank histograms for the two years (Fig. 15) for the eastern U.S. were computed to evaluate differences between forecasted and expected probability density functions (*pdf*s). Excessive population of the extreme ranks characterizes day 1 (Fig. 15a) and day 2 (Fig. 15b) of both summers. By day 2 (Fig. 15b), the over population obtains its extreme value as judged from the sum of the absolute difference of rank frequencies. The distributions at day 2 smoothly evolve to the ones shown for days 5 (Fig. 15c) and 10 (Fig. 15d), where the extremes are only slightly over populated. The excessive population of the end ranks, which is sustained to varying degrees throughout the forecast period, most likely reflects a model error whose cause is unknown at this time. Significance testing (Appendix) suggests that the distributions for the summers do not differ significantly at the 5% level after day 2, whereas each summer distribution differs at the 0.01% level relative to a uniform rank for all projections (Table 1).

These results suggest that the inclusion of evolved of singular vectors (*Barkmeijer et al*. 1999) had statistically significant impact on EPS performance for 24 h precipitation during the warm season to day 1 or 2. The improvement is related to a reduction of outlying verifications not captured by the EPS, as noted earlier by *Barkmeijer et al.* (1999) for different fields. No significant impact occurred after day 2. Of course, these conclusions rest on the assumption that the two summers were comparably difficult to forecast.

Brier Skill Scores	Summer 1997 Summer 1998	Winter 1997-1998 Winter 1998-1999
1 mm	1	1
10 mm	INSIG	1-4
20 mm	1-2	1-4
50 mm	1	2-4

ROC Areas	Summer 1997 Summer 1998	Winter 1997-1998 Winter 1998-1999
1 mm	INSIG	1-2
10 mm	INSIG	1
20 mm	8-9	1
50 mm	7	INSIG

Rank histograms	Summer 1997 Summer 1998	Winter 1997-1998 Winter 1998-1999
	1-2	1-4

Rank histograms	Summer versus Uniform Distribution	Winter versus Uniform Distribution
	1-10	1-10

TABLE 1: Forecast days when differences between accuracy measures for the indicated thresholds are statistically significant at the 5% confidence level based on non-parametric resampling tests (Appendix). The "INSIG" label indicates differences are not significant at any forecast time.

5.2 Evolved singular vectors and stochastic physics during the cool season

A spatially and temporally correlated stochastic term in the equation for temperature tendency was introduced in the operational EPS on 21 October 1998. *Buizza et al.* (1999c) show that the inclusion of stochastic components to sample sub-grid scale variability can improve EPS performance, particularly during the early stages of the forecast. The impact of this change, in conjunction with evolved singular vectors, on cool season precipitation can be assessed by comparing winter 1997-98 and winter 1998-99, where the winter season is defined as initial times during the period 1 November to 31 January, the latest initial time for which data are available.

Evolutions of the BSS and area are shown in Fig. 16 and 17 for the two cool seasons. In contrast to the two summers without stochastic physics, comparison of the two winter periods suggests that the implementation of stochastic physics and evolved singular vectors may have yielded a substantial improvement in accuracy and skill, especially in the short-range. The BSS values for days 1-3 typically run 0.2 to 0.4 units higher during the 1998-99 season. BSS differences are significant at the 5% level to day 1 for 1 mm and to day 4 for 10 to 50 mm (Table 1). Particularly noteworthy is the jump in skill for the 10 and 20 mm thresholds. For example, 20 mm exhibits no skill at day 1 in 1997-98, but remains skillful past day 5 in 1998-99. Decomposition of the Brier Score (Figs. 18 and 19) for 1-20 mm amounts indicates that both the reliability and resolution terms are improved in 1998-99 for the short-range. Even the BSS for 50 mm suggests somewhat of an improvement in the sense that the large dip in skill at day 2 is at least elevated to the level of climatology. Results for the ROC curves generally mimic those for the BSS, although ROC differences are significant for a shorter period, only to day 2 for 1 mm and to day 1 for 10 mm and 20 mm (Table 1). ROC areas for the 1-20 mm categories during 1998-99 are above 0.9 to day 2, and remain above the 0.8 level to day 5. Scores for the prior winter run typically run ~0.05 units lower from day 1 to 3. Differences between ROC areas for the heaviest category are not statistically significant at any time.

As before, rank histograms were used to evaluate differences in *pdf*s. Just like situation during the warm seasons, the cool seasons also exhibit signs of under dispersion at day 1 (Fig. 20a) and day 2 (Fig. 20b). The preponderance of over population of the extreme ranks is considerably lower for 1998-99, however. Note that frequency at day 1 in the lowest (highest) rank is 0.23 (0.16) for 1997-98, but it is only 0.07 (0.11) for 1998-99. The rank distributions for days 1-2 are clearly much closer to uniform in 1998-99 than in 1997-98. In fact, the three times reduction of the leftmost rank, relative to a much smaller reduction in the rightmost rank, is consistent with a decrease in a wet bias in the EPS. Note that both distributions are much closer to uniform by day 5 (Fig. 20c), and closer still by day 10 (Fig. 20d). The significance testing (Table 1) reveals that the rank distributions for the two winters differ significantly at the 5.0% level to day 4, but they are insignificant afterwards. Despite differences between forecast ranks and a uniform distribution steadily decreasing to day 10 for both years, both winter distributions still differ significantly at the 5.0% level at all projections relative to a uniform rank.

The comparative results for the two winters indicate that the changes to the EPS between the 1997-98 and 1998-99 winters markedly improved short-range to early medium-range (day 1 to day 4) performance for PQPF, again assuming the two winters were of comparable difficulty to forecast. It is not possible to isolate with total confidence the relative impact of evolved singular vectors opposed to stochastic physics. Nonetheless, the small differences between the 1997 and 1998 summer seasons, a period during which only evolved singular were implemented in the EPS, suggest that the wintertime improvements are most likely due to the inclusion of stochastic physics. These results are in agreement with the earlier results of *Buizza et al.* (1999c), who compared ensembles with and without stochastic physics but for a set of only 14 cases. It is not known at this time whether the implementation of stochastic physics produces comparable gains in skill during the warm season.

5.3 Interannual variability of root-mean-square error growth and predictability

As noted in the prior two subsections, an obstacle to determining whether an apparent improvement in the EPS performance is mainly due to system changes is the fact that results are for different years. For example, the atmosphere could have been inherently more predictable during the La Niña winter of 1998-99 than El Niño winter of 1997-98, which in turn could lead to better performance, independent of EPS changes. The uncertainty terms BS_{unc} for thresholds of 10 mm, 20 mm and 50 mm are much smaller for winter 1998-99 than the prior winter (Figs. 18 and 19). This indicates a lower frequency of occurrence for these thresholds during winter 1998-99 and a possible reduction in the forecast difficulty.

As a way to check the intrinsic predictability of the two years, the evolution of the RMSE of precipitation amount for the eastern U.S. was computed for the high-resolution (T_L319L31) deterministic forecasts, the highest resolution ECMWF model that was run during the period of study. The high-resolution forecasts start from an unperturbed initial analysis and contain fixed physics. Thus, differences between the two years can not be due to evolved singular vectors or stochastic physics. The ECMWF deterministic T_L319L31 model and data assimilation system (e.g. *Gerard and Saunders* 1999) did experience several changes over the course of study though, so the versions in 1998 or early 1999 were not the same as the ones in 1997. It is believed that the impact on mid-latitude QPF from the changes to the T_L319L31 analysis-forecast system is probably much smaller than the impact from the changes to the T_L159L31 EPS, evolved singular vectors and stochastic physics. So if the difference between the high-resolution deterministic forecasts is small, but the difference between the two ensembles is large, it follows that the EPS most likely improved because of the system changes, and not differences in predictability.

The RMSE curves are shown in Fig. 21 for the two summers and winters. The respective curves in a season appear very similar to each other. The two curves for summer differ by less than 10% out to 4 days, and by more than 10% afterwards. Differences between the winter curves tend to run somewhat larger for the short-range. The curves agree to within ~10% out to day 5, except at day 3 when they differ by ~20%. Differences beyond day 5 are ~5%. The statistical significance of the differences between the two curves for the same season was tested by permutation techniques (Appendix) for each forecast projection. Only summertime projections for days 5-10 exhibited differences significant at 5% confidence level. Days 1-4 for the summer and all projections for the winter had an average significance level of ~50%, the level expected due to chance, which suggests that differences in predictability characteristics between years are neither exceptionally small or large.

From this analysis, we conclude that the overall predictability properties of the different years were comparable to day 10 in winter and to day 4 in summer. In view of the fact that differences between the RMSE precipitation curves are far from significant for projections during which statistically significant differences exist for several accuracy metrics, we conclude that the improvements in ensemble performance most likely reflect changes in the EPS.

6. Case studies of heavy precipitation events

Analysis in the sections 4 and 5 confirms that the ECMWF EPS produces skillful PQPFs for 24-h accumulations, with predictions during the winter and for smaller thresholds generally possessing skill for a longer time. It also shows that skillful PQPFs for heavy precipitation events, episodes of 50 mm or greater that generally have the severest societal impacts (*Fritsch et al.* 1998), remain elusive, especially during the warm season. In this section, EPS performance is examined from the perspective of synoptic case studies, with the goal being a subjective identification of any obvious systematic error characteristics of PQPFs. Such documentation can aid operational forecasters and model developers by helping them identify error signatures, such as systematic phase shifts, that can not be inferred from single value accuracy measures such as the BSS and ROC curves.

Potential target cases were identified from inspection of daily time-series of the RMSE of the ensemble mean for day 4 and day 5 prognoses. Figure 22 shows time-series of the RMSE, averaged for the day 4 and day 5 forecasts valid for the same date, for the 1998 warm and 1998-99 cool seasons. The figure also shows daily time-series of the spatial RMS for the RFC verification grid, where the spatial RMS is defined as the square root of the sum over the entire grid of the precipitation squared. Thus, the RMS statistic is related to sum of area-averaged precipitation plus deviations from the spatial mean. Besides large day-to-day fluctuations in forecast accuracy, the two time-series exhibit a strong positive correlation (0.92 summer, 0.93 winter), indicative of large forecast errors being associated with heavy precipitation events. The ratio of the two curves averages ~1.2 during winter and ~1.0 during summer and fluctuates surprisingly little with date, which indicates that absolute errors in quantity are comparable to the observed precipitation amounts almost every forecast day.

A verifying date was considered for detailed examination if its RMSE was the largest within a 15-day period centered on the valid time. The plus/minus one-week window ensured that the same synoptic event (e.g. the same frontal wave cyclone) was not selected more than once. Five events from the cool season (verifying dates between 6 NOV 98 to 3 FEB 99) and eight events from the warm season (verifying dates between 6 MAY 98 to 4 OCT 98) with maximum RMSE values above 10 mm were selected for closer inspection. The dates are listed in Table 2. Interestingly, the same cases appear if the RPSS is used instead of RMSE.

Daily PQPFs from the EPS, RFC-based verifying analyses, and 500 mb height and surface analyses (NOAA 1998, 1999) were consulted for the 13 cases. What follows is a brief summary of error characteristics for the cool and warm seasons, based on our subjective synopsis of individual maps.

Summer Events	Winter Events
5 JUN 98	13 NOV 98
14 JUN 98	12 DEC 98
27 JUN 98	02 JAN 99
26 JUL 98	23 JAN 99
23 AUG 98	30 JAN 99
4 SEP 98	
12 SEP 98	
29 SEP 98	

Table 2: Dates for case studies of events with a regions of verifying precipitation greater than 50 mm in 24 h and relative large forecast errors at days 4 and 5.

6.1 Cool season characteristics

Examination of the upper-level and surface charts reveals that all five cases are associated with baroclinic cyclones, frontal zones and a strong northward surge of moist air from the Gulf of Mexico. A review of composite radar maps shows that the heaviest rainfall occurred near regions with cumulus convection. To illustrate common features of EPS performance, we will focus discussion on the middle case of 2 JAN 99.

The verifying 24 h rainfall analyses for 1200 UTC 2 JAN 99 (Fig. 23a) shows an area of > 50 mm centered over the Louisiana-Arkansas border, with the region of > 20 mm extending along the western Mississippi River Valley. Probabilities from the EPS for > 50 mm at projections of 1 day, 3 days and 5 days, all valid 1200 UTC 2 JAN 99, are shown in Figs 23b-23d. The 1-day forecast (Fig. 23b) reveals some overlap of regions of non-zero probabilities and the occurrence of 50 mm or greater rainfall, but the correspondence is so small that the 1-day BSS is slightly negative. The four other cases also exhibit some overlap, with the BSS varying from marginal skill to no skill. In four of five cases, the center of maximum probabilities is displaced by more than 250 km to the north (nearest quadrant) of maximum analyzed precipitation. The 3-day forecast (Fig. 23c) also shows some overlap and a northward displacement in three cases, with smaller maximum probabilities than at 1 day. The BSS is marginally skillful at day 3. The 5-day probability field (Fig. 23d) provides little hint of the impending heavy precipitation event, with a modest area of >1% likelihood of 50 mm about 500 km to the northeast of the observed event. All five cases at day 5 show a similar northward displacement error. The tendency for northward displacement of maximum predicted probabilities relative to verification suggests a possible phasing error in EPS forecasts of heavy wintertime precipitation. Note that such systematic errors can not be inferred from scalar accuracy measures such as the BSS, RPSS and ROC areas.

6.2 Warm season characteristics

The four earliest cases in Table 2, like the wintertime events, are associated with baroclinic cyclones and frontal zones. The last four entries in Table 2, however, are associated with the landfall of tropical storm Charley (23 Aug 98) and hurricanes Earle (4 SEP 98), Frances (12 SEP 98) and Georges (29 SEP 98). For this reason, we will discuss two cases to illustrate common features of EPS performance.

The case of 05 JUN 98 exemplifies, as much as any case can, behavior associated with baroclinic disturbances during the warm season. The verifying analysis for 05 JUN 98 (Fig. 24a) places a 100 mm maximum over western Tennessee, with amounts above 20 mm extending from eastern Missouri into northern Georgia. The EPS forecasts a maximum (14%) likelihood of > 50 mm over central Illinois at day 1 (Fig. 24b), ~200 km to the north of the verifying area. No ensemble members at day 3 (Fig. 24c) or day 5 (Fig. 24d) predict precipitation above 50 mm over the eastern U.S. Poor or erratic performance epitomizes the other three cases. Only the event of 14 JUN 98, a case of oceanic cyclogenesis off the coast of New England that is arguably rather winter-like in terms of the strength of baroclinic forcing, shows any skill at day 1 or day 3. Moreover, its performance is very erratic between consecutive forecast cycles, with skillful BSS values at days 1 and 3 but unskillful values at day 2 and after day 3.

On the other hand, short-range EPS forecasts for the tropical cyclones appear to fare somewhat better. The landfall of hurricane Georges was a particularly successful forecast. The verifying analysis for 1200 UTC 29 September 1998 (Fig. 25a) shows amounts > 50 mm along the central coast of the Gulf of Mexico. The 1-day probability forecast (Fig. 25b) is close to perfect, with a BSS of 0.91. The 3-day forecast (Fig. 25c), though sharp, is only marginally skillful (BSS=0.05) because the center of high probabilities is displaced ~200 km to the west-southwest of the observations. The 5-day forecast (Fig. 25d) shows only one grid box in a couple of ensemble members, improperly placed, with precipitation greater than 50 mm. The other two cases of hurricane landfall also show appreciable skill at day 1 (BSS of 0.48 for Earle and 0.17 for Frances) for amounts greater than 50 mm, but they are unskillful at longer projections. The landfall of the weaker surface circulation, tropical storm Charley over southern Texas, is unskillful even at 1 day.

7. **Discussion**

It is of interest to compare the performance of the EPS against the performance of the NCEP global ensemble (*Eckel and Walters* 1998). Differences in experimental designs between the two studies preclude detailed quantitative comparisons. Nonetheless, a qualitative comparison of Brier Skill Scores is of value and reveals some similar error characteristics. To facilitate comparison, verification procedures for the EPS output were modified to be more consistent with the methodology of *Eckel and Walters* (1998). Because the NCEP ensemble would likely benefit from verification on a coarser 2.5° x 2.5° grid since the less predictable scales are sampled less thoroughly (*Islam et al.* 1993), the EPS output and corresponding verifying analyses were spatially smoothed by a 9-point low-pass filter (*Shapiro* 1970, eq. 31) with a half-power cutoff of $4\Delta x$ to simulate a coarser resolution system more consistent with a 2.5° x 2.5° grid. To eliminate any possible benefit from a larger ensemble size, only 11 members from the EPS (the unperturbed forecast plus the first 5 pairs of twin perturbed runs), the same ensemble size as the NCEP system, were verified. Lastly, EPS scores were averaged for the cool and warm seasons, as in *Eckel and Walters* (1998), while their results (given in inches) were linearly interpolated to metric thresholds of 2 mm, 10 mm and 20 mm. Remaining differences in experimental designs between the two studies (e.g. verification periods, forecast cycles) are beyond our control.

Practical limits of skill for PQPF for the ECMWF EPS and NCEP ensemble can be estimated from the evolution of the BSS (Fig 26). Both ensemble systems are skillful relative to long-term climatology. Limits for the ECMWF EPS appear to be ~2 days longer for 2 mm and for 10 mm thresholds, while both systems exhibit similar skill for 20 mm amounts. The important question of how combining output from the two ensemble systems to form a multi-analyses, multi-model medium-range ensemble (e.g. *Evans et al.* 2000) might improve PQPF skill warrants evaluation, is beyond the scope of this paper.

Both ensemble systems also show an odd dip in the BSS at ~2 days for 20 mm per day. Curiously, the NCEP ensemble also uses a type of dynamically conditioned perturbations, bred modes (*Toth and Kalnay* 1993), that samples growing modes during the analysis cycle. The fact that the two ensemble systems both use dynamically conditioned initial perturbations suggests that the peculiar behavior may be related in part to the perturbation strategy. This hypothesis could be readily tested by running the EPS with a non-dynamical perturbation scheme, for example a Monte Carlo approach based on an assimilation of perturbed observations (e.g. *Houtekamer and Mitchell* 1998), and then checking the forecasts for signs of the same behavior. It would also be of interest to examine EPS performance with singular vector perturbations for a norm, optimization time and resolution more appropriate for parameters associated with grid-resolvable stratiform precipitation and grid-resolvable conditioning of the pre-convective environment.

The dip in short-range skill might also be related to a deficiency in the variance associated with high frequency, mesoscale precipitation events in the EPS. Atmospheric fields invariably possess “red” spectra (e.g. *Gilman* 1963), with variances that tend to decease with higher frequencies. This raises the possibility that a short-range forecast might be more degraded by insufficient high frequency activity than a medium-range or extended-range forecast. For example, phenomena strongly tied to the diurnal cycle and shorter-lived intermittent fluctuations, such as cumulus convection, would be sampled multiple cycles over the first 2-3 days of the forecast. On the other hand, the more energetic synoptic- (periods 5-7 days) and planetary-scale waves (periods 10 days and longer) would be sampled only over a partial cycle. It follows that in the absence of strong multiplication noise (e.g. *Sardeshmukh et al.* 2000), a short-range forecast would be relatively more affected by high-frequency transience than a medium-range forecast because too little time passes for the more powerful slower transients to be fully sampled. But if the model variance associated with the high frequencies is deficient, then a short-range forecast would suffer from under dispersion because of the intrinsic relationship between ensemble spread and the model's variance (e.g. *Simmons et al.* 1995), even if the synoptic and planetary waves possess proper variance. This under dispersion in turn would lead to reduced skill for an ensemble PQPF. In view of this reasoning, it is not surprising that the implementation of stochastic physics in the EPS, with a temporal correlation scale of 6 hours that increases high-frequency variance (*Buizza et al.* 1999c), would improve probabilistic skill in the short-range but not the medium-range.

As noted earlier, the EPS produces skillful PQPFs in terms of Brier scores and ROC areas to 10 days for the lowest thresholds. Probabilistic skill to 10 days, taken at face value, might seem contrary to prior results that indicate the synoptic-scales lose predictability by 5-7 days for such fields as 500 mb height or sea-level pressure (e.g. *Lorenz* 1969, *Lorenz* 1982, *Baumhefner* 1984, *Nutter et al.* 1998). Several factors must be weighted when comparing our results with prior estimates of predictability limits, however. First, our results are for a 24-h accumulation interval on a 1.25° by 1.25° grid. Predictability limits would certainly decrease if a shorter accumulation period was used or if verification was performed against rain gauges or on a finer grid, since the less predictable smaller scales would be sampled more completely (*Islam et al.* 1993). Another consideration, related to the first factor, is the role that the planetary-scale waves play in modulating synoptic-scale, baroclinic-wave activity. The planetary-waves are predictable past 10 days for 500 mb height (*Nutter et al.* 1998), which suggests that preferred regions for enhanced storm-track activity are likely predictable past 10 days too, even if the individual synoptic features are not. A final, important point to consider is the difference between probabilistic skill and deterministic skill. Deterministic skill for a perfect model is lost once the ensemble variance exceeds the climatological variance. Probabilistic skill for a perfect model is not lost until the model’s entire attractor is filled and the ensemble’s pdf can no longer be distinguished from the model’s climatology (*Palmer* 1995), a situation that occurs after deterministic skill is lost. The time until loss of skill obviously becomes shorter with an imperfect model, but even with an imperfect model deterministic skill would still be lost prior to probabilistic skill except for scales that are nonlinearly saturated at the initial time.

Reliability diagrams for the EPS indicate a systematic tendency to over forecast the probability of precipitation for all thresholds and projections. Over prediction appears to be a ubiquitous feature of ensemble forecast systems (e.g. Fig. 6 of *Hamill and Colucci* 1998, Fig. 12 of *Eckel and Walters* 1998). As noted earlier, reliability curves for the EPS show a monotonic relationship between forecast probability and conditional probability of occurrence. This consistency suggests that the reliability of EPS precipitation forecasts could be easily enhanced through calibration by statistical post-processing techniques. Prior calibrations of ensemble output and precipitation forecasts by methods besides multiple linear regression have lead to significant increases in skill (e.g. *Hamill and Colucci* 1998, *Mullen et al.* 1998, *Eckel and Walters* 1998, Hall et al. 1999). For example, the calibration of the NCEP ensemble extended the useful limit of probabilistic skill by one day (*Eckel and Walters* 1998). Their success with the NCEP ensemble suggests that a quick way to improve PQPF's from the EPS is through statistical calibration. Of course, future refinements in data assimilation, the generation of initial perturbations or the model formulation would likely increase the reliability of the raw PQPF's, which in turn would likely improve statistical forecasts based on EPS predictors.

Consistently skillful PQPF's of heavy precipitation events by the EPS, or by present-day ensemble systems in general (e.g. *Hamill and Colucci* 1998, *Du et al.* 1997, *Eckel and Walters* 1998), remain elusive. The EPS does exhibit some forecasts of remarkable skill for the short-range (e.g. 14 June 1998, 29 Sept. 1998), while others are highly unskillful (e.g. 23 August 1998, 20 January 1999) or lack consistency between consecutive forecast cycles (e.g. 14 June 1998). Faced with such day-to-day volatility, operational forecasters might lack confidence in ensemble guidance, even if long-term verification statistics indicate useful skill. One major challenge, among many, confronting the research community in its pursuit of consistently skillful PQPF's is the rarity of heavy precipitation events. Sample grid point frequencies for 50 mm episodes, which can be inferred from the uncertainty term of the Brier Score decompositions, are ~0.4% during summer and ~0.9% during winter for 50 mm episodes defined on a 1.25° by 1.25° grid. Precipitation forecasts at adjacent grid points are not independent, however, which reduces even further the effective number of degrees of freedom (dof's). If the impact of spatial interdependence on dof's is considered, then the sample frequency for independent grid points with heavy precipitation is probably closer to the order of 0.1% for grid points with heavy precipitation. Thus, only ~100 of the grid points that verified above 50 mm, of ~35,000 total grid points in a full 5-month season (summer or winter), can likely be considered independent for our sample. An effective sample size of order 100 observed occurrences is far too small to generate robust estimates for reliability diagrams and ROC curves. The paucity of independent events in a single season, along with large degree of interannual variability (*cf.* uncertainty terms of Figs 18 and 19), points to the need to evaluate regional PQPF performance for many (~100) heavy rain events over several seasons, for several distinct regions. Once performance is confidently assessed and the leading error sources are identified, research can be specifically tailored to refine EPS forecasts of heavy precipitation. A byproduct of improved forecasts for heavy events should be even more skillful forecasts for lower thresholds. If centers of high forecast probabilities for heavy thresholds coincided more closely with occurrences, then centers of high probabilities for the low thresholds would likely improve too since they tend to be erroneously displaced in the same sense. In fact, the strong correlation between ensemble mean precipitation and probabilities for low thresholds for the EPS (results not shown) suggests that increasing the accuracy for low thresholds might improve the forecast positions of heavy episodes too.

Another uncertainty of this study, independent of the analysis-model system, involves the verifying analyses. Only RFC gauge data with a crude quality check were used as input for a very simple objective analysis scheme. There is certainly the question of how representative an average based on as few as five, heterogeneously distributed rain gauges is for precipitation in a 1.25° by 1.25° grid box. It seems likely that our analyses are more uncertain in the summer than winter, and in regions with heavy rainfall or convective precipitation. In fact, precipitation analyses for select times and regions likely contain as much error and uncertainty as precipitation forecasts themselves. The need for more accurate precipitation analyses that blend observations from a variety of instruments (i.e. gauges, radar, satellite) and that resolve the spatial and temporal mesoscale variability of rainfall is widely recognized (*Fritsch et al.* 1998). Consistent with the philosophy of ensemble forecasts, we add that a thorough analysis of precipitation should also include an estimate of its uncertainty. Future verification studies of ensemble PQPF that use blended analyses, complete with confidence bounds, are clearly desirable. It is also of interest to verify ensemble PQPF for spatial and temporal scales of greater interest to the hydrological community, such as watersheds for accumulation periods other than 24 h (*Fritsch et al.* 1998, *Buizza et al.* 1999a).

8. Summary

In this paper, the performance of the ECMWF Ensemble Prediction System was documented for probabilistic forecasts of 24-h precipitation over the eastern United States. Verification was performed against analyses based on rain gauge data from the River Forecast Centers of NOAA. Skill was measured relative to long-term climatic frequencies. We summarize main findings and highlight important conclusions below.

- Precipitation is more predictable during the winter than the summer, presumably because the synoptic forcing is stronger and convection is less prevalent.

- The EPS produces significantly skillful forecasts to past one week for a threshold of 1 mm in both seasons. The long, ~10 day limit for PQPF's of 1 mm is likely related to the lengthy 24 h accumulation period and large grid size of 1.25° latitude by 1.25° longitude. Accuracy decreases as the threshold increases, until forecasts of 50 mm are not skillful at 1 day for either season.

- The implementation of evolved singular vectors in the EPS appears to have minimal impact on skill during the summer. Conversely, the addition of stochastic processes in the EPS appears to raise short-range skill for thresholds up to 20 mm.

Research is underway to examine the impact of increased resolution on EPS medium-range forecasts of heavy precipitation. Results will be reported in due course.

Acknowlegdements

The authors thank Dr. T. Hamill of NCAR and two anonymous referees for their insightful reviews. We also thank Dr. A. Simmons of ECMWF for his useful comments on an early version of the manuscript. M. Baldwin of the NOAA Storm Prediction Center provided access to the RFC data. M. Wandishin of the NOAA National Severe Storm Laboratory supplied the long-term climatic frequencies. The first author was supported by ATM9612487, ONR N00014-98-0551 and ONR N00014-99-1-0181.

APPENDIX

A.1 Reliability, Brier Score and Brier Skill Score

Reliability (e.g. *Wilks* 1995, pg 236*ff*) measures the correspondence between the forecast probability y_i and the conditional probability of the event occurring given it was forecast $\overline{o}_i = p(o_i|y_i)$ where *i* denotes a discrete probability interval and the overbar a subsample average for category *i*. A reliability diagram can be constructed by plotting $\overline{o}_i$ as a function of y_i for all values of *i*. The curve for a perfectly calibrated forecast lies along the 1:1 diagonal.

The Brier score is an accuracy measure that is related to the reliability diagram. The Brier score is zero for a perfect forecast, and its upper bound is one. The Brier score can be decomposed into the sum of three terms related to reliability, resolution and uncertainty (e.g. *Wilks* 1995, pg 260-262)

$$BS = BS_{rel} - BS_{res} + BS_{unc}$$

where

$$BS_{rel} = \frac{1}{N}\sum_{i=1}^{I} N_i\left[y_i - \overline{o}_i\right]^2 \quad ,$$

$$BS_{res} = \frac{1}{N}\sum_{i=1}^{I} N_i\left[\overline{o}_i - \overline{o}\right]^2 \quad ,$$

$$BS_{unc} = \overline{o}\left[1 - \overline{o}\right],$$

and *N* is the total number of forecast/event pairs and $\overline{o}$ is the sample climatology. The quantities with *i*-subscripts denote subsample values of N_i, y_i $\overline{o}_i$ for discrete categories in 10% intervals from 0% to 100%, so *I*=10 for this choice.

The reliability term BS_{rel} summarizes the degree of agreement between forecast probabilities and observed frequency of occurrence for the event. The BS_{rel} term is zero for a forecast system with perfect reliability. The resolution term BS_{res} measures the degree to which the forecast frequencies differ from the sample climatology. The BS_{res} component is relatively large, and thus contributes to a smaller Brier Score, for forecasts that sort observations into categories with substantially different relative frequencies than the overall sample climatology. The uncertainty term BS_{unc} depends only on the overall sample climatology, and thus it is independent of the forecast system. The BS_{unc} term is large when the overall sample climatology is close to 50%, and forecasting the event is inherently more uncertain.

The Brier skill score is defined as the percentage improvement of Brier score with respect to climatology,

$$BSS = \frac{BS_{cli} - BS}{BS_{cli}},$$

where BS_{cli} is the Brier score for a forecast based on climatology, and climatology is computed from the Summary of the Day data. If $Bs_{cli} - Bs_{unc}$, then the BSS can be written in terms of reliability and resolution terms as

$$BSS = \frac{BS_{cli} - BS}{BS_{cli}} = \frac{BS_{res} - BS_{rel}}{BS_{unc}},$$

where the last term represents the skill with respect to sample climatology. The $BBS = 1$ for a perfect forecast, and $BSS \leq 0$ for a no-skill forecast.

Four thresholds of 24 h accumulated precipitation greater than amount P_p mm/24h are considered, with P_p=1, 10, 20 and 50. For each threshold, probabilities are defined as the ratio of the number of ensemble members forecasting the event over the total number of ensemble members.

A.2 Ranked Probability Score and Ranked Probability Skill Score

The Ranked Probability Score (RPS; *Epstein* 1969, *Murphy* 1971) is a generalization of the Brier score to multicategory, probabilistic forecasts. The RPS measures the square error between a forecast cumulative frequency distribution (*cfd*) and an observed *cfd* , summed over J mutually exclusive, collectively exhaustive categories. A RPS of zero denotes a perfect forecast, that is the forecast probability is 1 in the correct category. The worst possible score is J-1. The same four thresholds for the Brier score are used to define categories for the RPS, which gives J=5 categories with lower bounds of zero and P_p=1, 10, 20 and 50.

The Ranked Probability Skill Score (RPSS) is analogous with the BSS,

$$RPSS = \frac{RPS_{cli} - RPS}{RPS_{cli}},$$

where is the RPS_{cli} for a forecast based on climatology. Again, $RPSS = 1$ for a perfect forecast, and $RPSS \leq 0$ for a no-skill forecast.

A.3 Signal Detection Theory

A four-member contingency table can be constructed for a deterministic forecast for any dichotomous event (e.g. 24 h precipitation greater than 1 mm) as shown in Table A1. The entries in the contingency table can be used to define a hit rate as H/(H+M) and the false alarm rate as FA/(FA+CR). The hit rate and false alarm rate can be plotted on a scatter diagram for any contingency table.

Signal detection theory is an extension of the contingency table concept to probabilistic forecasts. A probabilistic forecast can be used to construct a set of contingency tables by defining a “yes” forecast to correspond to a probability exceeding a threshold, then varying the threshold. For example, forecast values equal to or greater than $1/M$, where M is the number of ensemble members, are considered a “yes” forecast, and corresponding hit rates and false alarm rates are computed. Then the procedure is repeated for values equal to or greater than $2/M$, $3/M$, , $(M\text{-}1)/M$. The process yields a set of hit rates and false alarm rates that can be plotted on a scatter diagram.

These points, and the two default points of (0,0) and (1,1) that correspond to hit rate and false alarm rate for the 0% (never forecasting the event) and 100% (always forecasting the event) threshold, can be connected by line segments to define a curve, the Relative Operating Characteristic (ROC) curve. The area under the ROC curve can be used as a simple measure of forecast quality. An area of 1.0 corresponds to a perfect forecast system: 100% hit rate and 0% false alarm rate. A useless forecast lies along the 1:1 diagonal and has an area of 0.5, which corresponds to equal hit rates and false alarm rates, and hence no ability to discriminate a forecast event.

The area under ROC curve is qualitatively similar to resolution in the sense that it assesses the ability of the forecast system to discriminate between occurrences and non-occurrences. The ROC curve provides no information about reliability though, since it is based on stratification by observations.

A.4 Verification of Rank

For a hypothetical “perfect model” ensemble forecast system, one where all forecast errors are due to errors in the initial state and are randomly sampled, the value of the verifying observation *x* when pooled with *M* ensemble members that are sorted in ascending order $x_1,\ldots, x_M$ ($x_i < x_j$ for $i < j$) is equally likely to occur in each of M+1 ranks. Summing the rank over many independent realizations, a uniform distribution of the ranks is expected. A consistent ensemble system divides the probability density function (pdf) of a specific forecast into a uniform distribution:

$$p_i \equiv \mathbf{Pr}(x \le x_i) = \frac{2i-1}{2M}$$

where verification is equally probable in any rank. If the rank distribution is not uniform, it indicates that the ensemble system lacks internal consistency: the forecast model is not perfect and/or the selection of initial perturbations is non-random (*Anderson* 1996, *Hamill and Colucci* 1998).

If the distribution exhibits a greater percentage in the extreme ranks than expected, it indicates insufficient variability and the ensemble is termed “under dispersive”. Conversely, if the distribution exhibits a greater percentage in the center ranks than expected, it indicates excessive variability and the ensemble is deemed “over dispersive”. Note that it is possible for an ensemble to contain a greater than expected percentage in both the extreme and the center ranks, and thus exhibit characteristics of both over and under dispersion.

A.5 Significance Testing

The resampling procedure first pools together all dates for two years for the same forecast projection, under the assumption of no significant temporal correlation between 24-h precipitation forecasts from consecutive analyses cycles (*Hamill* 1999). Two seasons worth of days are then randomly selected from the pool, and respective BSS values, ROC areas and rank frequencies are calculated for the two artificial data sets. Absolute differences are then computed between the resampled BSS values or ROC areas, or between the resampled rank frequencies with a chi-square type statistic $\sum_{i=1}^{I}\left[p_f(i) - p_e(i)\right]^2 / p_e(i)$ where I = 52 is the total number of ranks, $p_f(i)$ is the probability for rank *i*, and $p_e(i)$ is the probability for the expected distribution, either for rank *i* for the prior year or for a uniform distribution (1/*I*). The randomization procedure is repeated 10,000 times, and an empirical distribution of the different values, that is a background *cfd*, is formed from the permutations.

The resulting arrays of artificial differences are sorted in ascending order, and the ranks of the actual differences relative to the empirically derived background *cfd* are determined to give a significance level. The 95% (99%) position of the *cfd* corresponds to the 5% (1%) significance level. The significance level is the probability of incorrectly rejecting the null hypothesis. Here the null hypothesis that the two means or two distributions are equal is tested. In this study, an *a priori* significance level of 5% is selected as an acceptable probable error in rejecting the null hypothesis by just a fortuitous random sampling. The procedure is repeated separately for all forecast projections for the winter and summer seasons.

The resampling procedure can also be easily altered to give estimates of 90% (or any level) confidence bounds for the accuracy measures. This is accomplished by only pooling together all dates for a particular year or season (e.g. all winter days). Days are then randomly selected from the pool, and respective BSS values or ROC areas are calculated for the artificial data sets. Th procedure is repeated 10,000 times. The resulting array of artificial scores is sorted in ascending order, and the 5% and 95% rank relative to the empirically derived background *cfd* are determined to give an estimate of the 90% confidence bound.

	Yes Forecast	Not Forecast
Yes Observed	H (Hits)	M (Misses)
Not Observed	FA (False Alarms)	CR (Correct Rejections)

Table A1: Contingency table needed to compute ROC curves.

REFERENCES

Anderson, J.L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, 9, 1518-1530.

Barkmeijer, J., Buizza, R., and T.N. Palmer, 1999: 3D-Var Hessian singular vectors and their potential use in the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.,* 125, 2333-2351.

Barnes, S.L., 1973: Mesoscale objective analysis using weighted time-series observations. NOAA Tech. Memo. ERL NSSL-62, National Severe Storms Laboratory, Norman, OK, 60 pp. [NTIS COM-73-10781.]

Baumhefner, D.P. 1984: *Analysis and forecast intercomparisons using the FGGE SOP1 data base*. NAS FGGE Workshop, Woods Hole, Massachusetts.

Brier, G. W. 1950: Verification of forecasts expressed in terms of probabilities. *Mon. Wea. Rev.,* 78, 1-3.

Brooks, H. E., M. S. Tracton, D. J. Stensrud, G. J. DiMego, and Z. Toth, 1995: Short-range ensemble forecasting: report from a workshop, 25-27 July 1994. *Bull. Amer. Met. Soc.,* 76, 1617-1624.

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999a: Probabilistic predictions of precipitation using the ECMWF Ensemble Prediction System. *Wea. Forecasting*, 14, 168-189.

Buizza, R., J. Barkmeijer, T.N. Palmer, and D.S. Richardson, 1999b: Current status and future developments of the ECMWF Ensemble Prediction System. *Meteorol. Appl*., 6, 1-14.

Buizza, R., M. Miller, and T.N. Palmer, 1999c: Stochastic simulation of model uncertainties. *Quart. J. Roy . Meteor. Soc.*, 125, 2887-2908.

Buizza, R., T. Petroliagis, T.N. Palmer, J. Barkmeijer, M. Hamrud, A. Hollingsworth, A Simmons, N. Wedi, 1998: Impact of model resolution and ensemble size on the performance of an ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, 124, 1935-1960.

Buizza, R., 1997a: Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, 125, 99-119.

Buizza, R., 1997b: Some aspects of the performance of the ECMWF Ensemble Prediction System. *Proc. Expert meeting on the use and development of the Ensemble Prediction System*, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading RG2 9AX, UK, 86 pp.

Carter, G.M., J.P. Dallavalle, and H.R. Glahn, 1989: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Wea. Forecasting*, 4, 401-412.

Doswell, C.A., III, 1987: The distinction between large-scale and mesoscale contribution to severe convection: A case study example. *Wea. Forecasting*, 2, 3-16.

Du, J., S. L. Mullen and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.*, 125, 2427-2459.

Eckel, F.A. and M.K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. Forecasting*, 13, 1132-1147.

Epstein, E.S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, 8, 985-987.

Evans, R. E., M. S. J. Harrison, and R. J. Graham, 2000: Joint medium-range ensembles from the U.K. Meteorological Office and ECMWF systems. *Mon. Wea. Rev.*, 128, in press.

Fritsch, J.M., R.A. Houze Jr, R. Adler, H. Bluestein, L. Bosart, J. Brown, F. Carr, C. Davis, R.H. Johnson, N. Junker, Y.-H. Kuo, S. Rutledge, J. Smith, Z. Toth, J.W. Wilson, E. Zipser, and D. Zrnic, 1998: Quantitative precipitation forecasting: report of the eighth prospectus development team, U.S. Weather Research Program. *Bull. Amer. Met. Soc.*, 79, 285-299.

Gerard, E., and R. Saunders, 1999: 4Dvar assimilation of SSM/I total column water vapour in the ECMWF model. *Quart. J. Roy . Meteor. Soc.*, 125, 3077-3101.

Gilman, D.L., F.H. Fuglister, and J.M. Mitchell, 1963: On the power spectrum of "red noise". *J. Atmos. Sci.*, 20, 182-184.

Hall, T., H.E. Brooks, and C.A. Doswell III, 1999: Precipitation forecasting using a neural network. *Wea. Forecasting*, in press.

Hamill, T.M., and S.J. Colucci, 1998: Evaluation of the Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, 126, 711-724.

Hamill, T.M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, 14, 155-167.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, 124, 1225-1242.

Houtekamer, P.L. and H.L. Mitchell 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.*, 126, 796-811.

Huffines, G.R. and R.E. Orville, 1999: Lightning ground flash density and thunderstorm duration in the continental United States: 1989-96. *J. Appl. Meteor.*, 38, 1013-1019.

Islam, S., R.L. Bras, and K.A. Emanuel, 1993: Predictability of mesoscale rainfall in the Tropics. *J. Appl. Meteor.,* 32, 297-310.

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

Lorenz, E.N. 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, 21, 289-307.

Lorenz, E. N., 1982: Atmospheric predictability experiments with a large numerical model. *Tellus*, 34, 505-513.

Mason, 1982: A model for assessment of weather forecasts. *Austr. Met. Mag.,* 30, 291-303.

Messinger, et al. 1996: Improvements in quantitative precipitation forecasts with the eta regional model at the National Centers for Environmental Prediction: the 48-km upgrade. *Bull. Amer. Meteor. Soc.,* 77, 2637-2649.

Molteni, F., R. Buizza, Palmer, T. N., and T. Petroliagis, 1996: The ECMWF ensemble prediction system: methodology and validation. *Quart. J. Roy. Meteorol. Soc*., 122, 73-119.

Mullen, S.L., M.M. Poulton, H.E. Brooks, T.M. Hamill, 1998: Post-processing of ETA/RSM ensemble precipitation forecasts by a neural network. Preprints, *1st Conference on Artificial Intelligence*, Phoenix AZ, Amer. Meteor. Soc., J31-J33.

Murphy, A. H., 1971: A note on the Ranked Probability Score. *J. Appl. Meteor*., 10, 155-156.

Murphy, A.H., and R.L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev*., 115, 1330-1338.

Murphy, A. H., 1991: Forecast verification: its complexity and dimensionality. *Mon. Wea. Rev*., 119, 1590-1601.

Nutter, P. A., S. L. Mullen and D. P. Baumhefner, 1998: The impact of initial condition uncertainty on numerical simulations of blocking. *Mon. Wea. Rev*., 126, 2482-2502.

NOAA, 1998 & 1999: *Daily Weather Maps*. [Available from Climate Prediction Center, World Weather Building, Room 811, Washington, D.C. 20223]

Palmer, T.N., F. Molteni, R. Mureau, R. Buizza, P. Chaplelet, an J. Tribbia, 1993: Ensemble prediction. *Proc. ECMWF Seminar on Validation of models over Europe: Vol. I*, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading RG2 9AX, UK, 286 pp.

Palmer, T.N., 1995: *Predictability of the Atmosphere and Oceans: From Days to Decades*. NATO Advanced Study Institute, ‘Decadal Climate Variability: Dynamics and Predictability’, Les Houches, February 1995.

Pauley, R., M. A. Rennick, and S. Swadley, 1996: Ensemble forecast product development at Fleet Numerical Meteorology and Oceanography Center. Preprints, *11th Conference on Numerical Weather Prediction*, Norfork, Virginia, Amer. Meteor. Soc., J61-J63.

Sanders, F. 1986: Trends in skill of Boston forecasts made at MIT 1966-84. *Bull. Amer. Meteor. Soc*., 67, 170-176.

Sardesmukh, P. D., G. P. Compo, and C. Penland, 2000: Changes of probability associated with El Nino. *J. Climate*, 13, in press.

Shapiro, R., 1970: Smoothing, filtering and boundary effects. *Reviews of Geophysics and Space Physics*, 8, 359-387.

Simmons, A. J., R. Mureau and T. Petroliagis, 1995: Error growth and estimates of predictability from the ECMWF forecasting system. *Quart. J. Roy. Meteorol. Soc*., 121, 1739-1772.

Stanski, H.R., L.J. Wilson, and W.R. Burrows, 1989: *Survey of common verification methods in meteorology*. World Weather Watch Tech. Rep. 8, WMO, Geneva, 114pp.

Stensrud, D.J., J.-W. Bao, and T.T. Warner, 1999: Using initial condition and model physics perturbations in short-range ensembles. *Mon. Wea. Rev*., 127, 433-446.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: the generation of perturbations. *Bull. Amer. Meteor Soc.*, 74, 2317-2330.

Toth, Z., E. Kalnay, S. Tracton, R. Wobus, and J. Irwin, 1997: A synoptic evaluation of the NCEP ensemble. *Wea. Forecasting*, 12, 140-153.

Toth, Z., Y. Zhu, T. Marchok, S. Tracton, and E. Kalnay, 1998: Verification of the NCEP global ensemble forecasts. Preprints, *12th Conference on Numerical Weather Prediction, Phoenix*, AZ, Amer. Met. Soc., 286-289.

Tracton, M.S. and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: practical aspects. *Wea. Forecasting*, 8, 379-398.

Tracton, M.S. and J. Du, 1998: Short-range ensemble forecast (SREF) at NCEP/EMC. Preprints, *12th Conference on Numerical Weather Prediction*, Phoenix AZ, Amer. Met. Soc. 269-272.

Wallace, J.M., 1975: Diurnal variations in precipitation and thunderstorm frequency over the conterminous United States. *Mon. Wea. Rev*., 103, 406-419.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, New York, 269-272, 467pp.

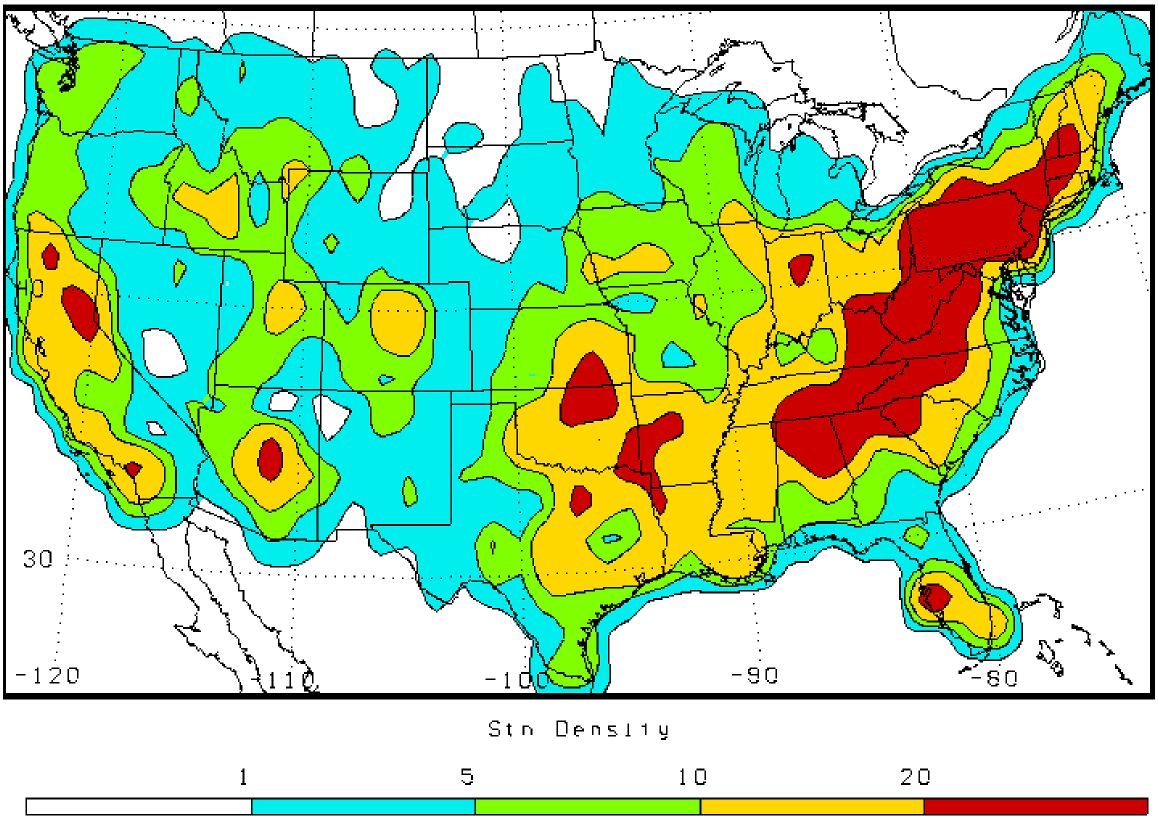


Fig. 1: Station density per 1.25° by 1.25° grid boxes for the RFC network, averaged over the entire study period.

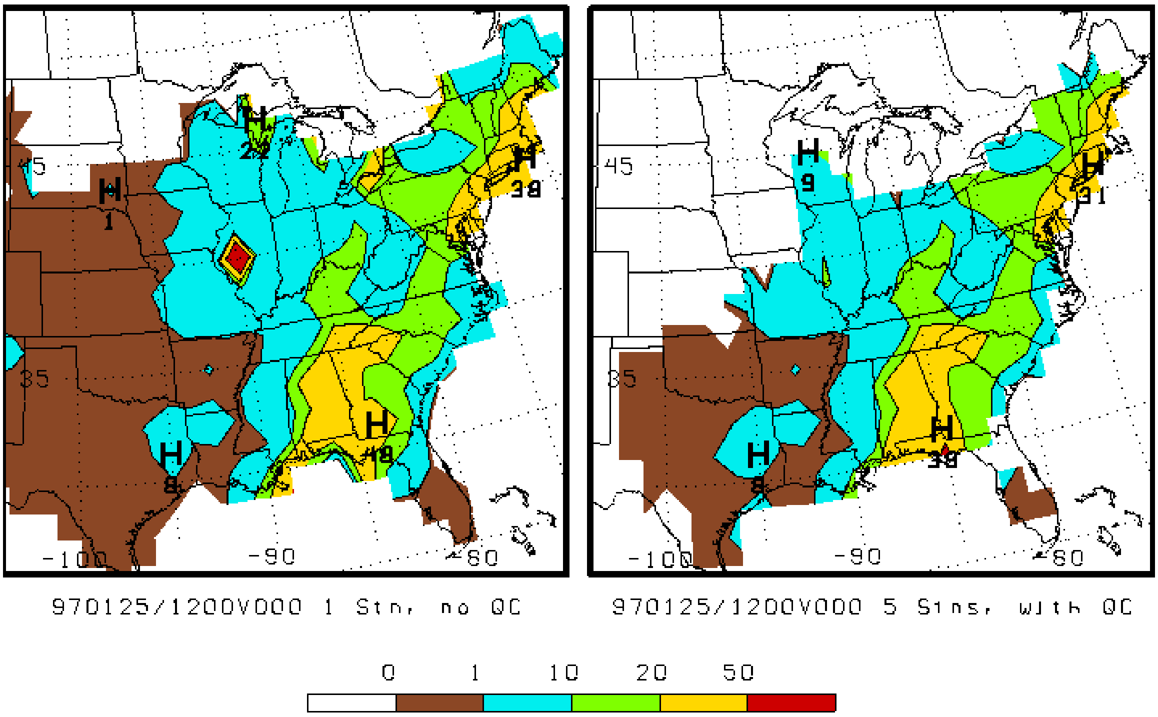


Fig. 2: Verifying 24 h precipitation analyses, valid 97011812, (a) without (left panel) and (b) with (right panel) criterion for minimum station density and data quality control. See text for details.

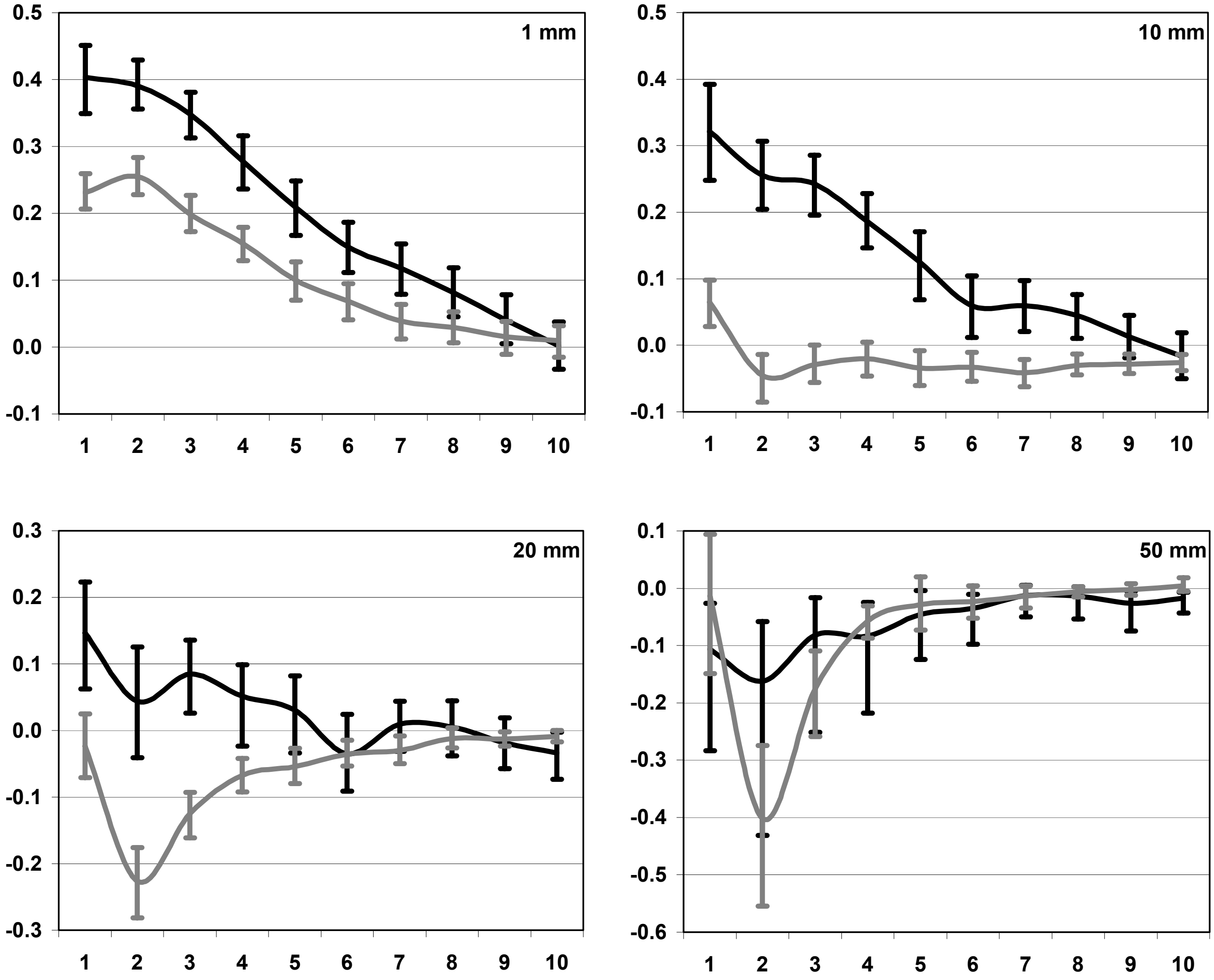


Fig. 3: Brier Skill Score, relative to long-term climatology, for the winter (black line) and summer (gray line) seasons. Upper left: 1 mm. Upper right: 10 mm. Lower left: 20 mm. Lower right: 50 mm. Error bars give 90% confidence bounds. Note different scales for the ordinates. See text and Appendix for details.

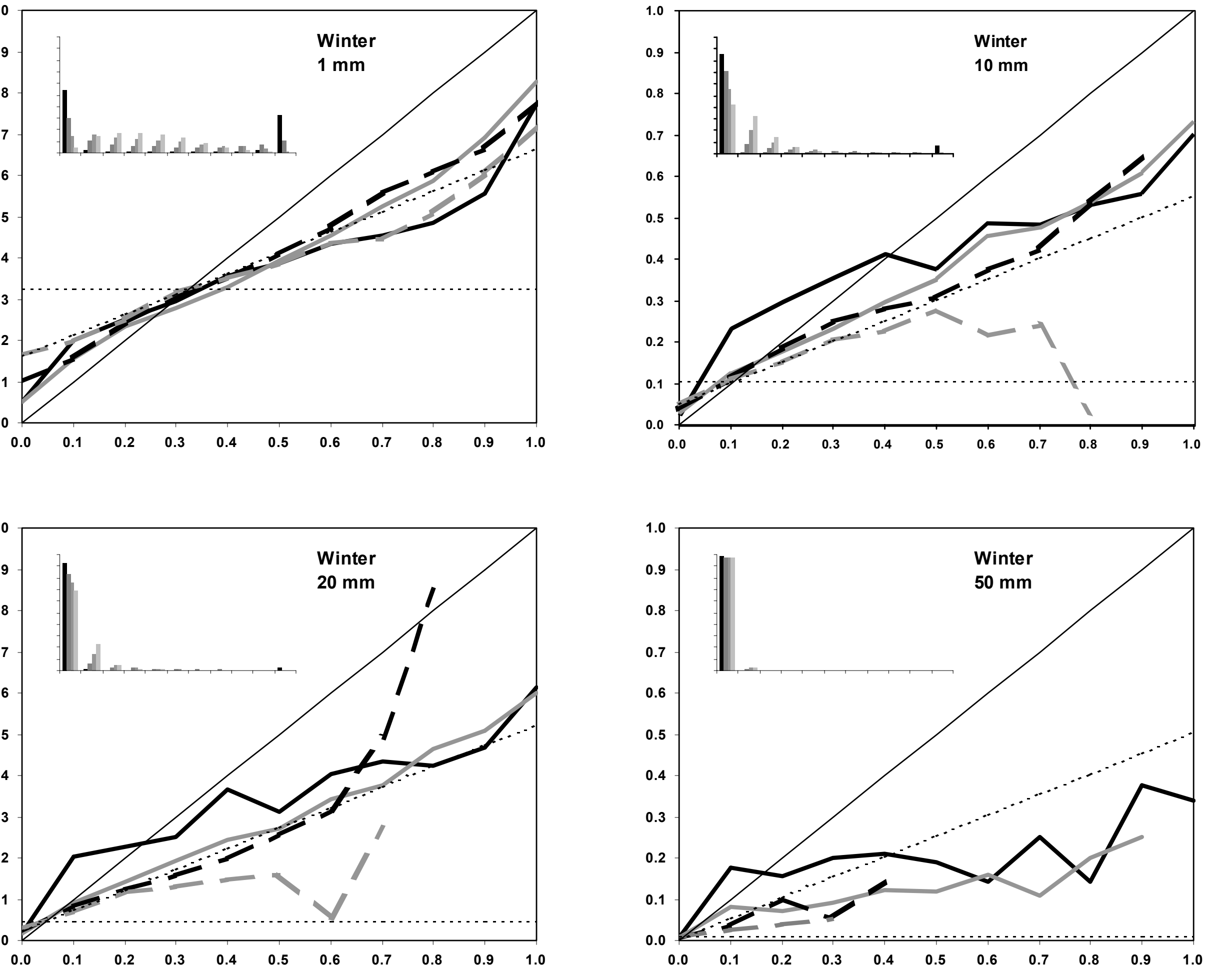


Fig. 4: Reliability diagrams for winter valid at Day 1 (solid black), Day 4 (solid gray), Day 7 (dashed black), Day 10 (dashed gray). Upper left: 1 mm. Upper right: 10 mm. Lower left: 20 mm. Lower right: 50 mm. The horizontal dashed line denotes the sample climatology, and the sloped dashed line denotes the no-skill line. Abscissa is forecast probability; ordinate is conditional probability of occurrence. The insert gives the frequency of the 11 forecast categories for the given threshold. Only points for forecast frequencies greater than 0.01% are plotted.

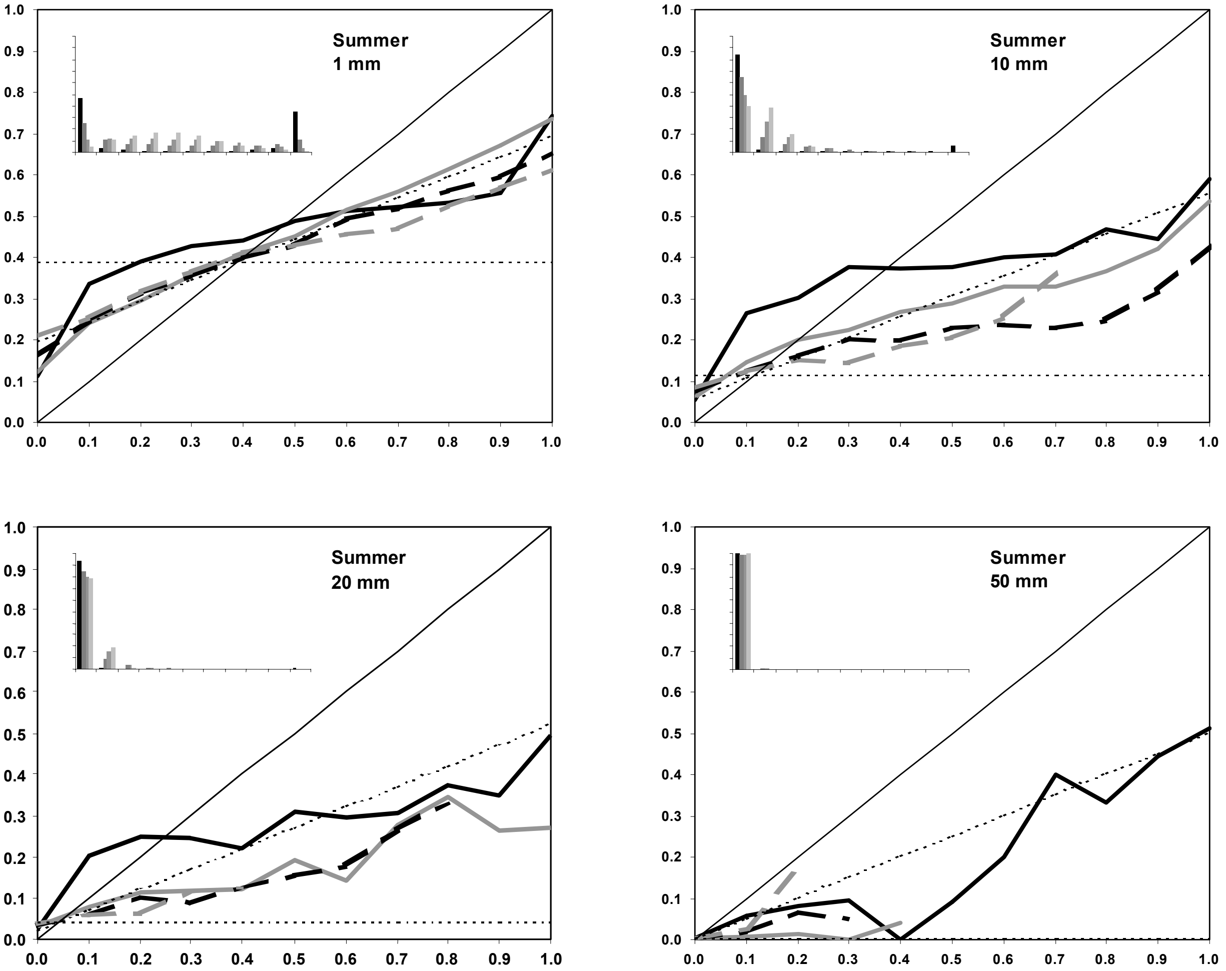


Fig. 5: As in Fig. 4, except for summer

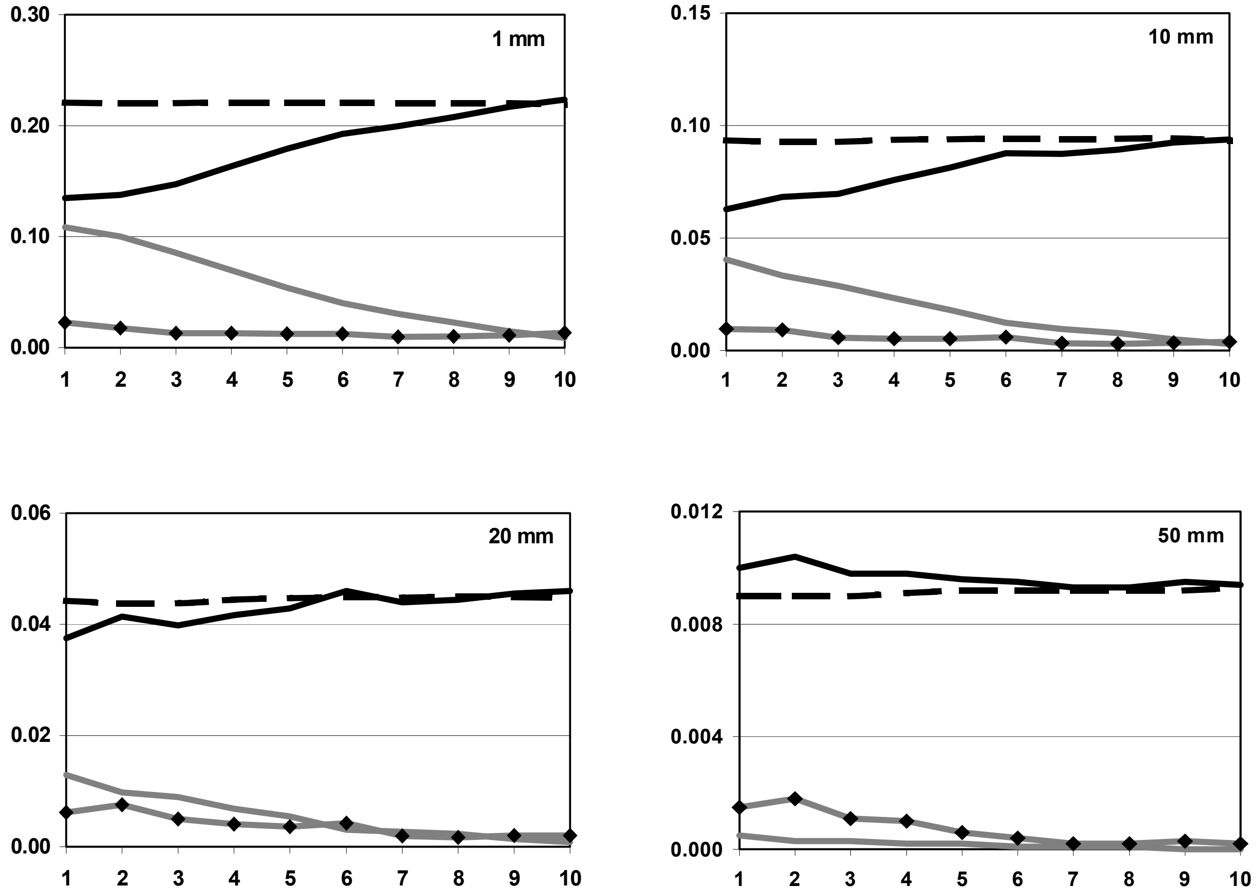


Fig. 6: Decomposition of the Brier Score (black line) into reliability (gray line with diamonds), resolution (gray line) and uncertainty (dashed line) terms for winter at the indicated thresholds. Forecast is skillful relative to sample climatology when the Brier Score (reliability term) exceeds the uncertainty (resolution) term.

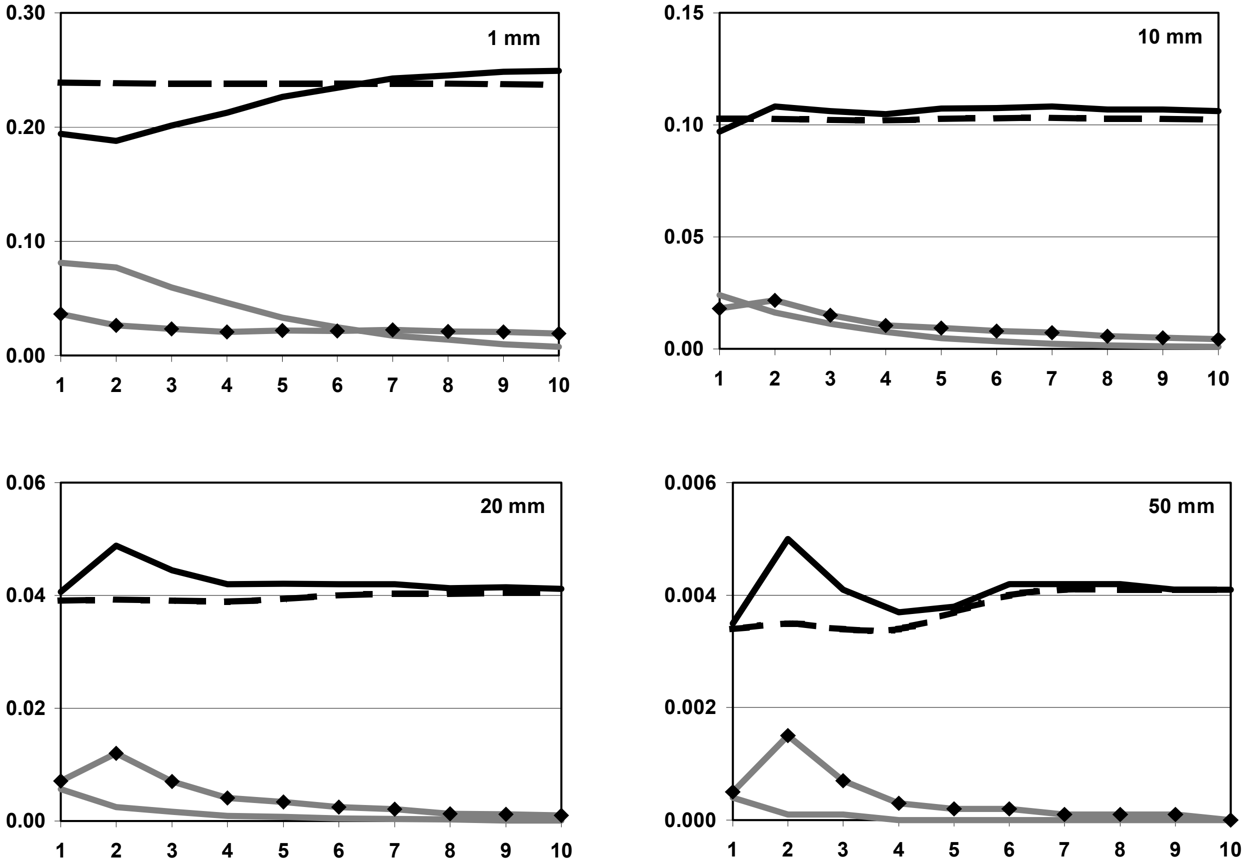


Fig. 7: As in Fig. 6, except for summer.

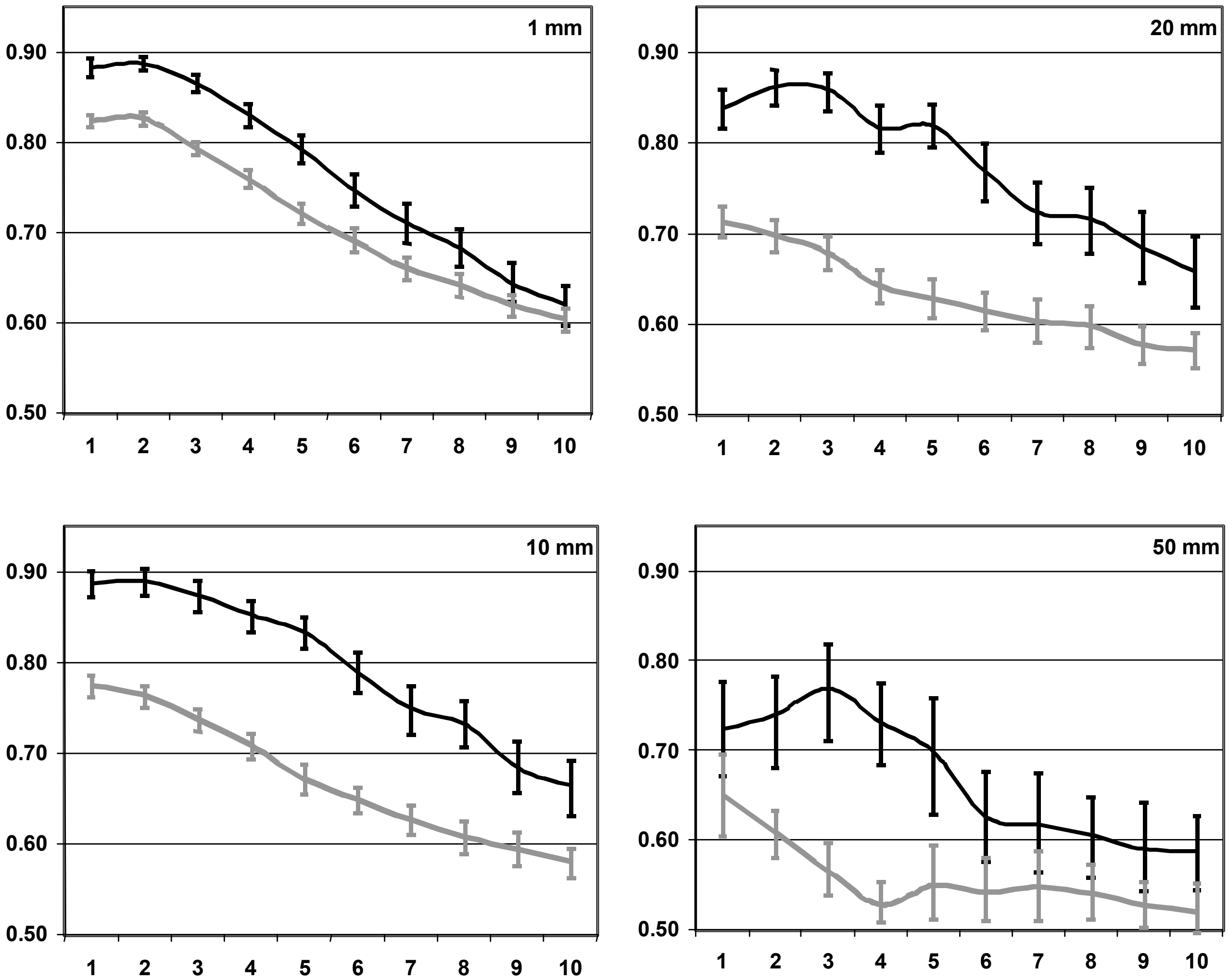


Fig. 8: Area under Relative Operating Characteristic curves for the winter (black line) and summer (gray line) seasons. Upper left: 1 mm. Upper right: 10 mm. Lower left: 20 mm. Lower right: 50 mm. Error bars give 90% confidence bounds. See text and Appendix for details.

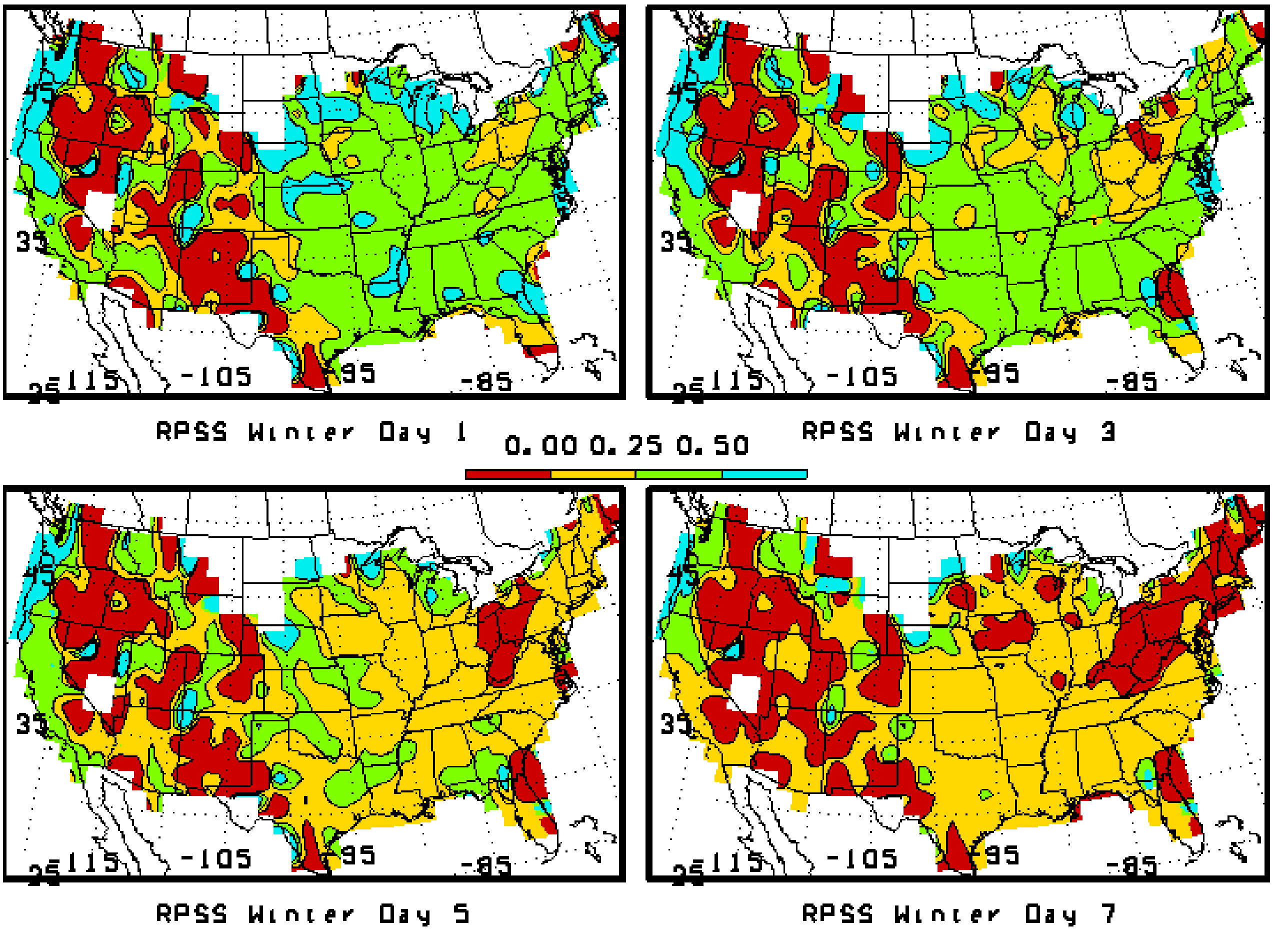


Fig. 9: Spatial distribution of the Ranked Probability Skill Score during the winter for 1 day (top left), 3 day (top right), 5 day (bottom left) and 7 day (bottom right) forecasts. No shading indicates regions not verified due to lack of observations. See text for details on calculation.

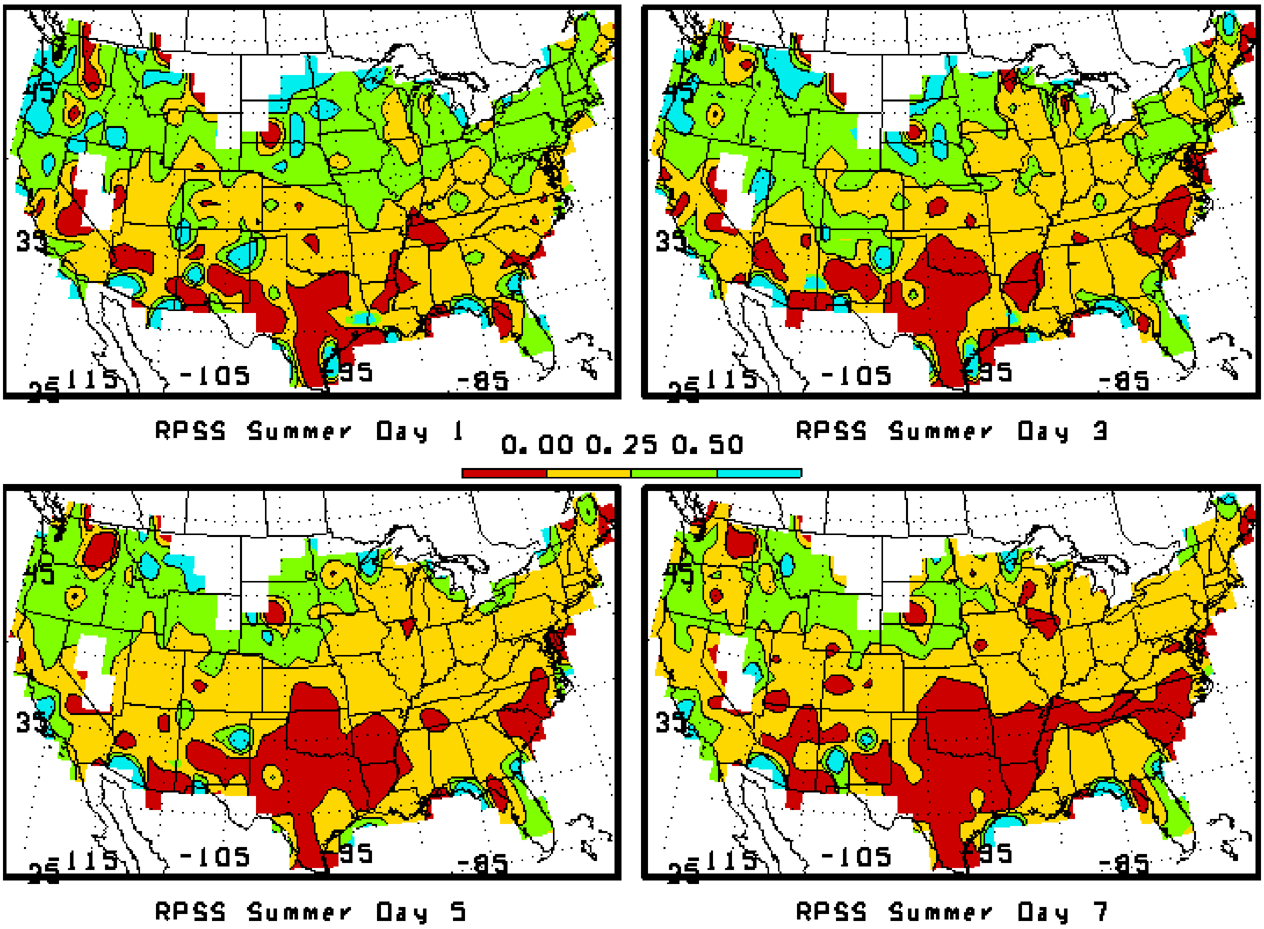
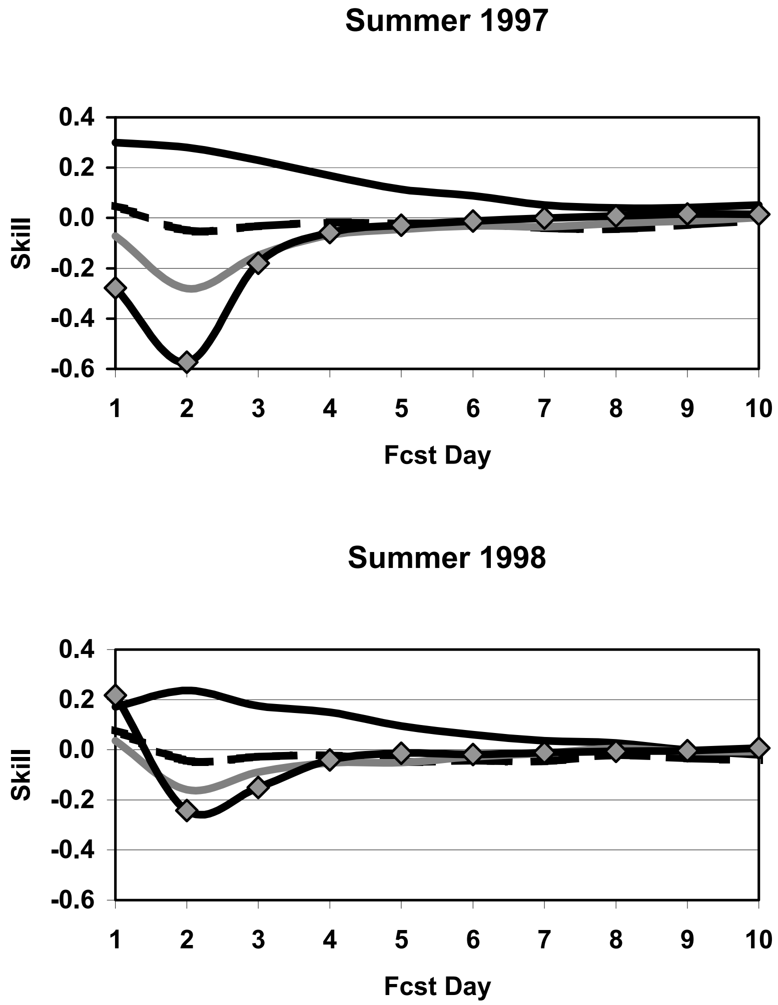
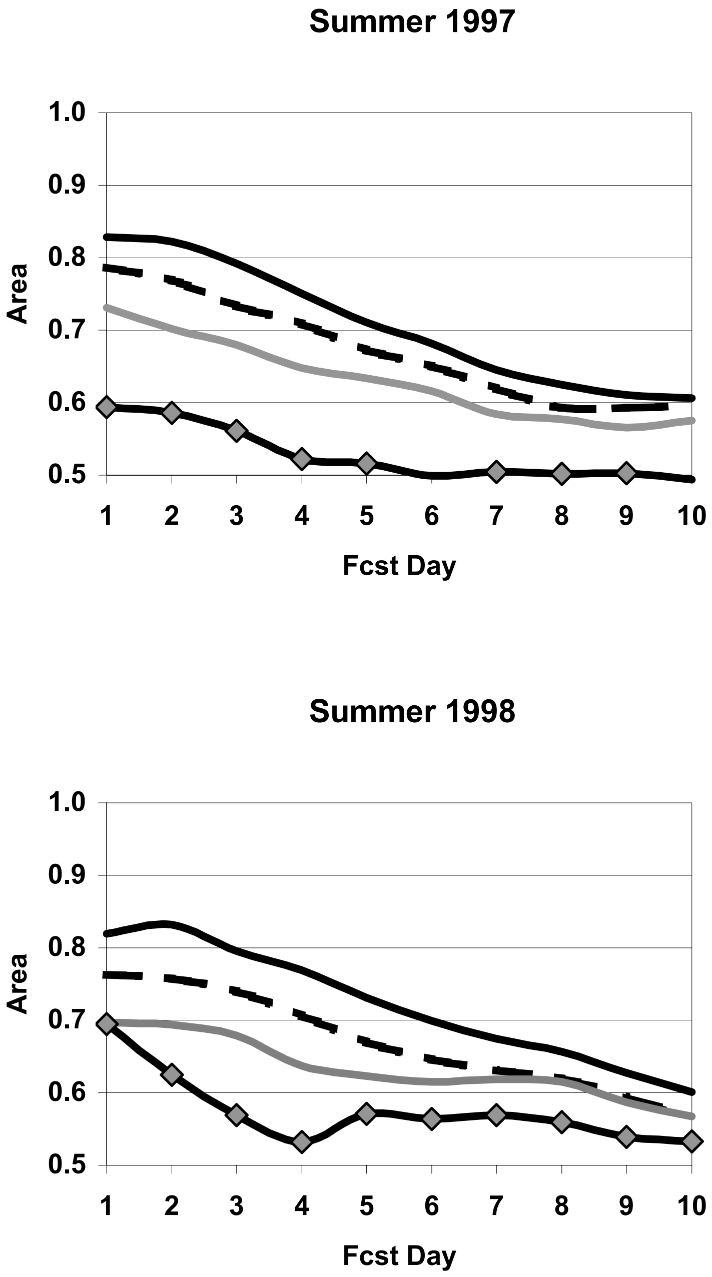


Fig. 10: As in Fig. 9, except for summer.


Summer 1997

Summer 1998

Fig. 11: Brier Skill Scores for summer 1997 and summer 1998. Thresholds are 1 mm (solid black), 10 mm (dashed black), 20 mm (solid gray), and 50 mm (black with diamonds).



Summer 1997

Summer 1998

Fig. 12: Area under Relative Operating Characteristic curve for summer 1997 and summer 1998. Thresholds are 1 mm (solid black), 10 mm (dashed black), 20 mm (solid gray), and 50 mm (black with diamonds).

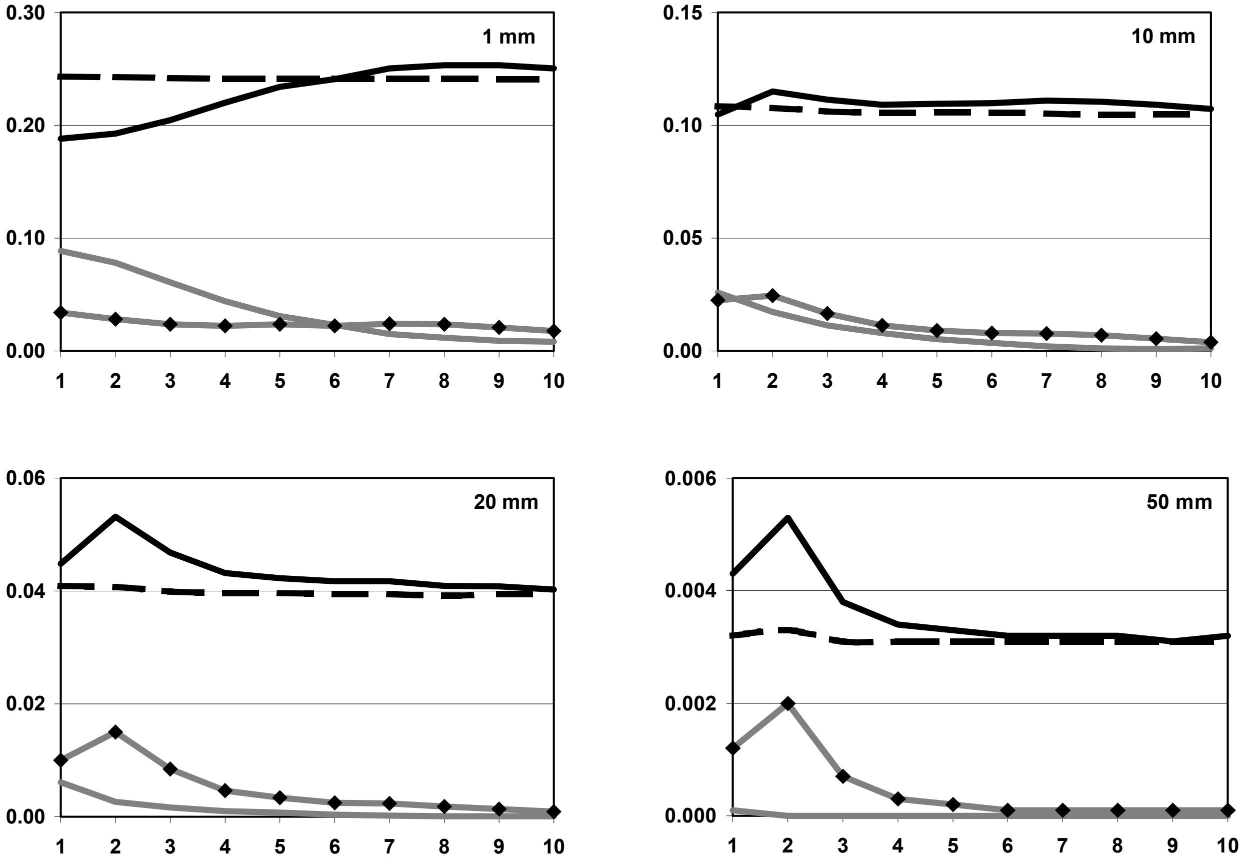


Fig. 13: As in Fig. 6, except for summer 1997

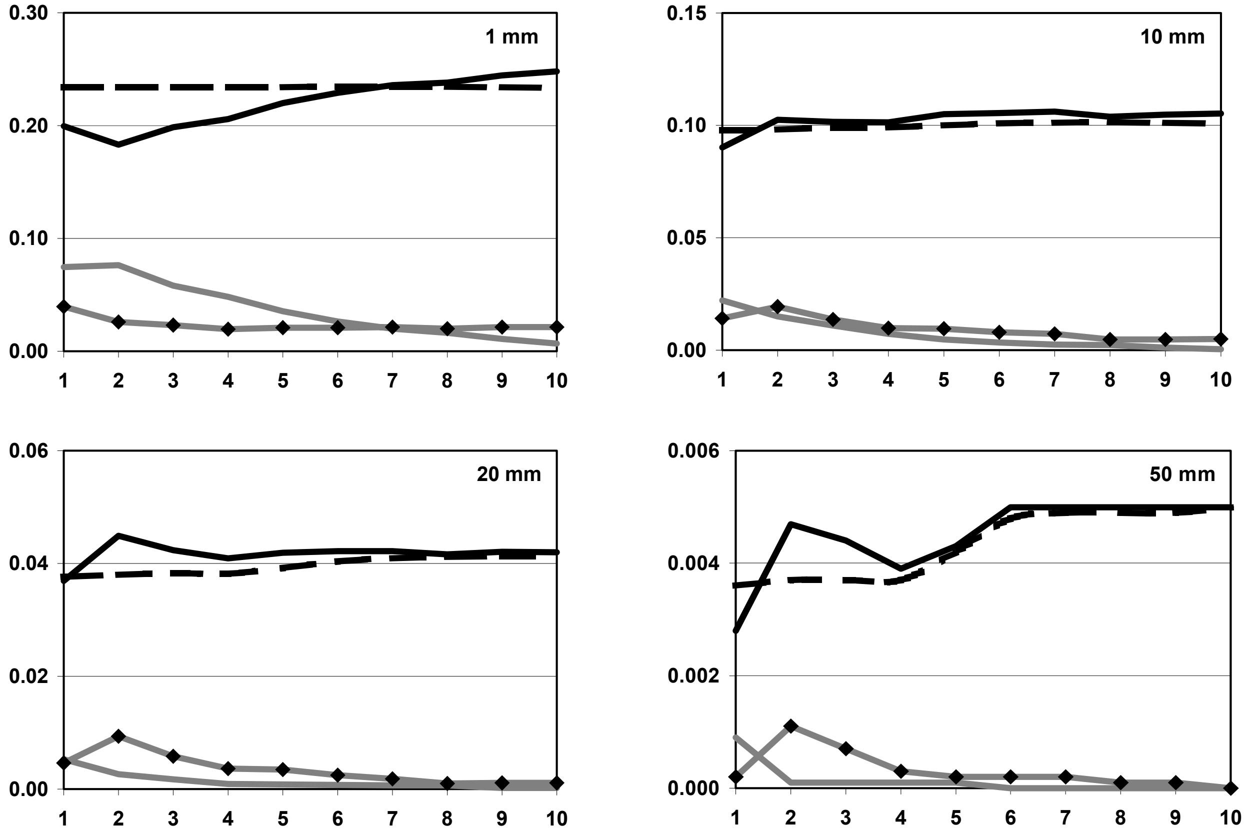


Fig. 14: As in Fig. 6, except for summer 1998.

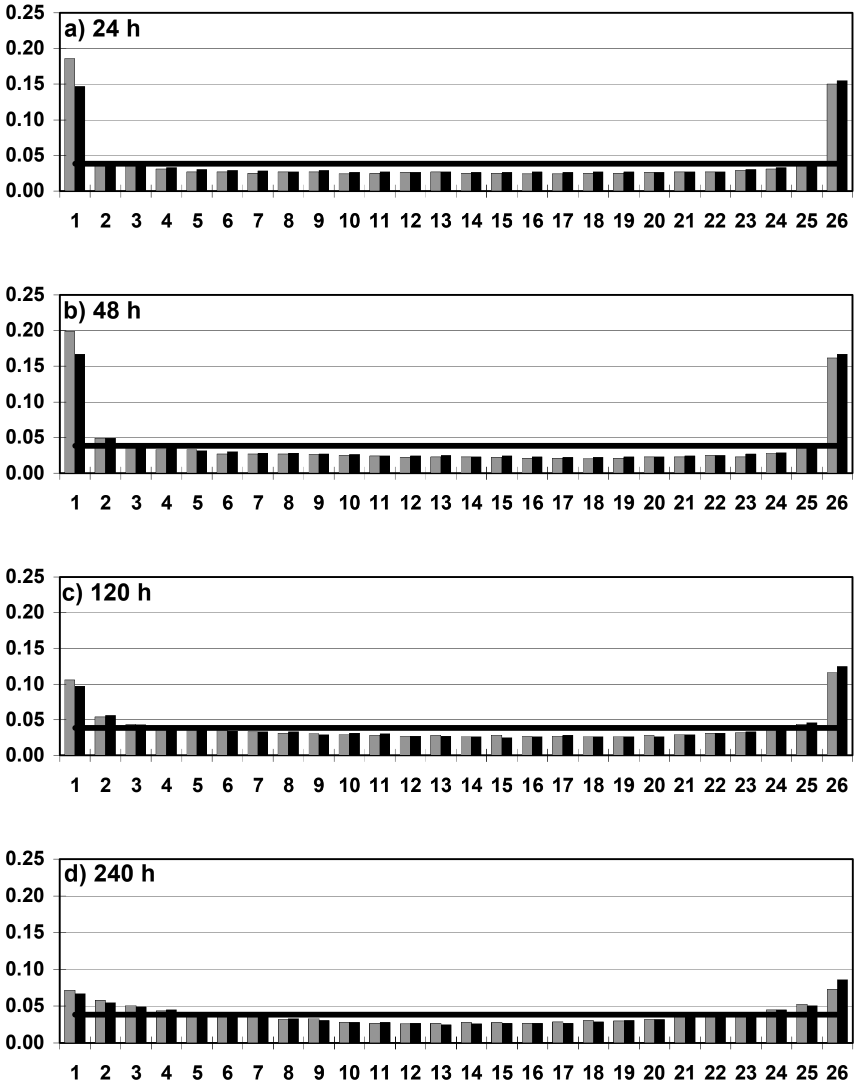
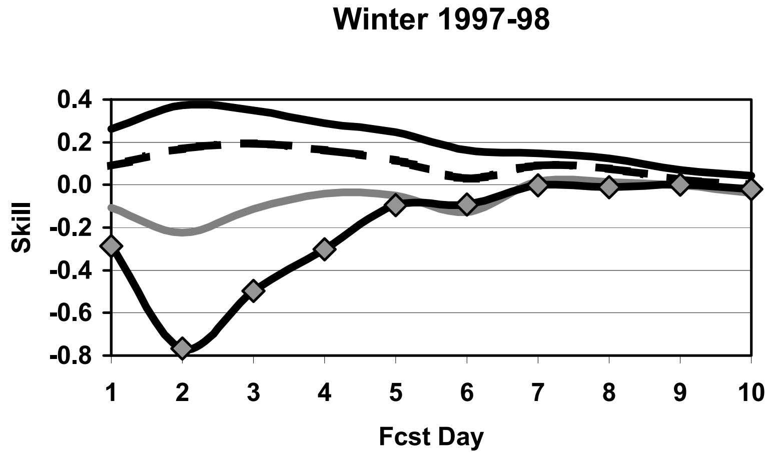


Fig. 15: Rank histograms for 24, h 48 h, 120 h and 240 h precipitation forecasts. Gray bars: summer 1997 prior to the implementation of evolved singular vectors. Black bars: summer 1998 after the implementation of evolved singular vectors. Heavy horizontal line denotes frequency for uniform rank distribution. For clarity, only 26 categories are plotted, where the abscissa index i shows the sum of rank (2i-1) + rank (2i). See text for details.



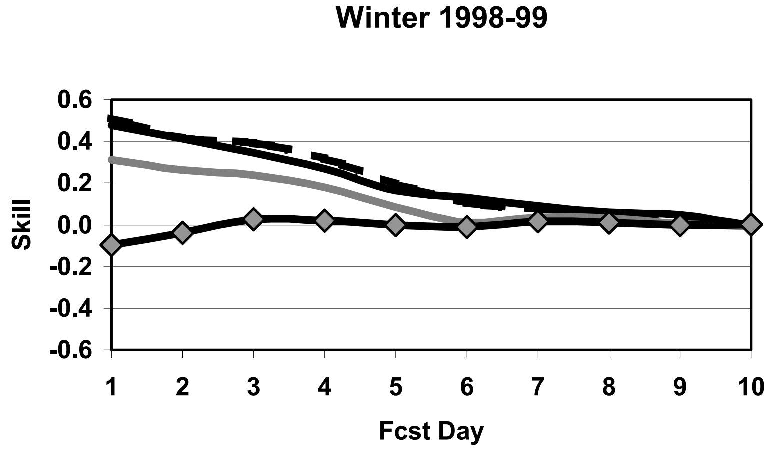
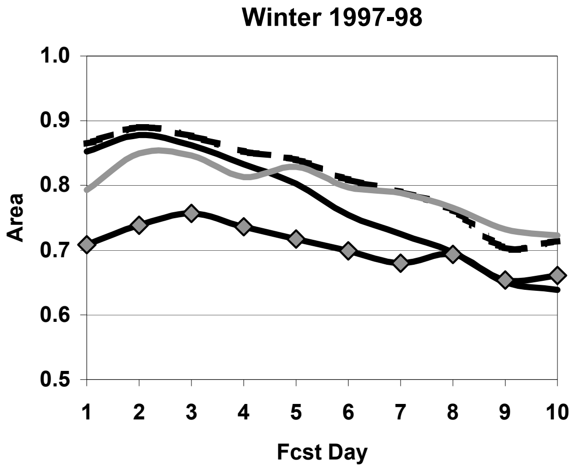


Fig. 16: Brier Skill Scores for winter 1997-98 and winter 1998-99. Note different scales for the ordinates. Thresholds are 1 mm (solid black), 10 mm (dashed black), 20 mm (solid gray), and 50 mm (black with diamonds).



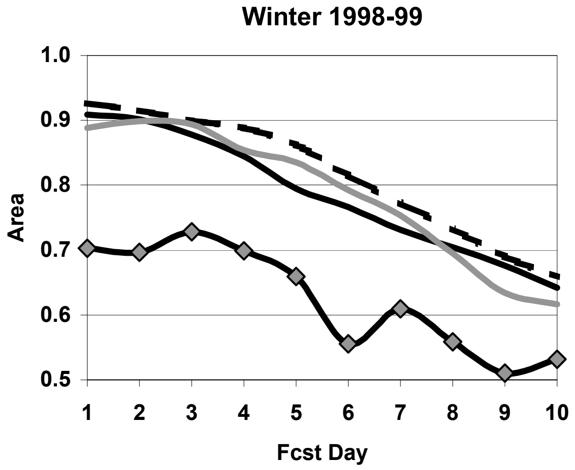


Fig. 17: Area under Relative Operating Characteristic curve for winter 1997-98 and winter 1998-99. Thresholds are 1 mm (solid black), 10 mm (dashed black), 20 mm (solid gray), and 50 mm (black with diamonds).

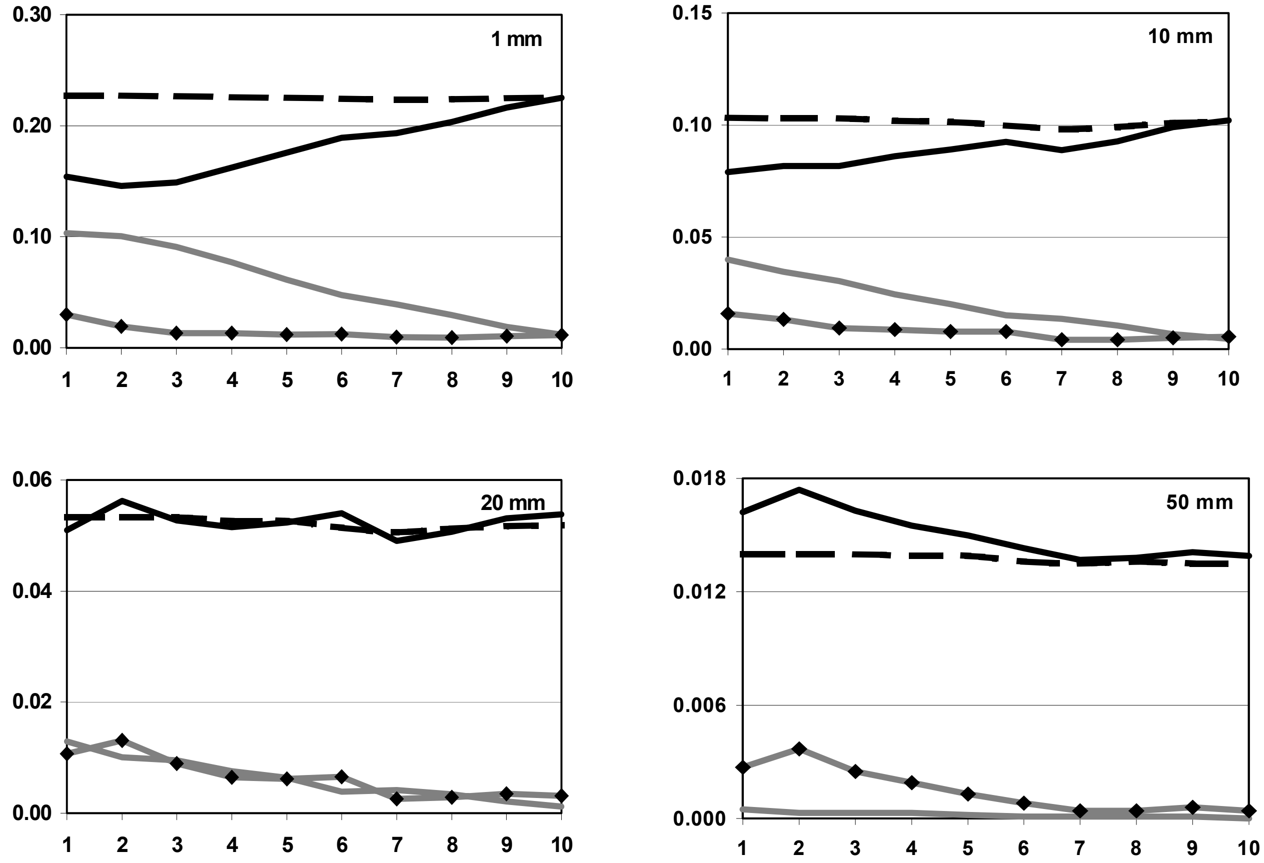


Fig. 18: As in Fig. 6, except for winter 1997-98.

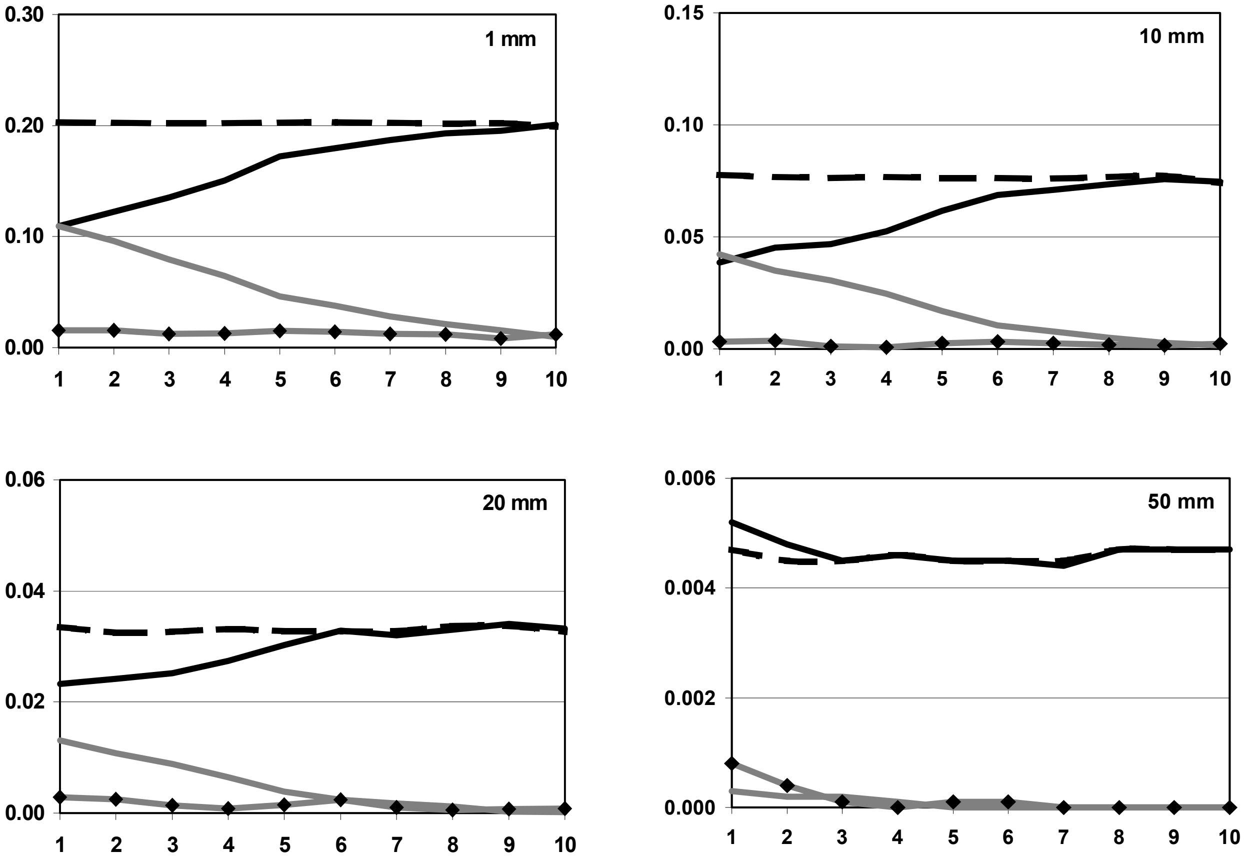


Fig. 19: As in Fig. 6, except for winter 1998-99.

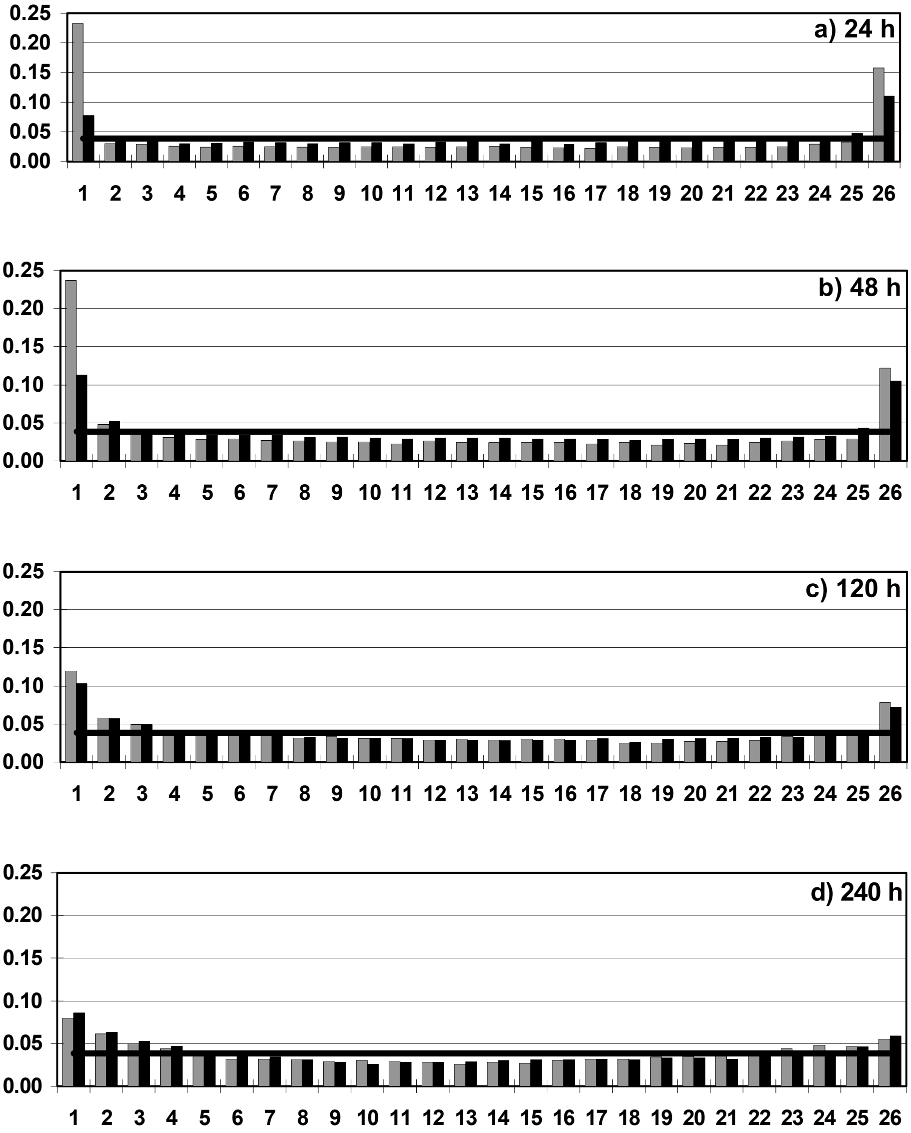


Fig. 20: Rank histograms for 24, 48, 120 and 240 h precipitation forecasts. Grey bars: winter 1997-98 prior to the implementation of evolved singular vectors and stochastic physics. Black bars: winter 1998-99 after implementation of evolved singular vectors and stochastic physics. Heavy horizontal line denotes frequency for uniform rank distribution. For clarity, only 26 categories are plotted, where the abscissa index i shows the sum of rank(2i-1) + rank(2i). See text for details.

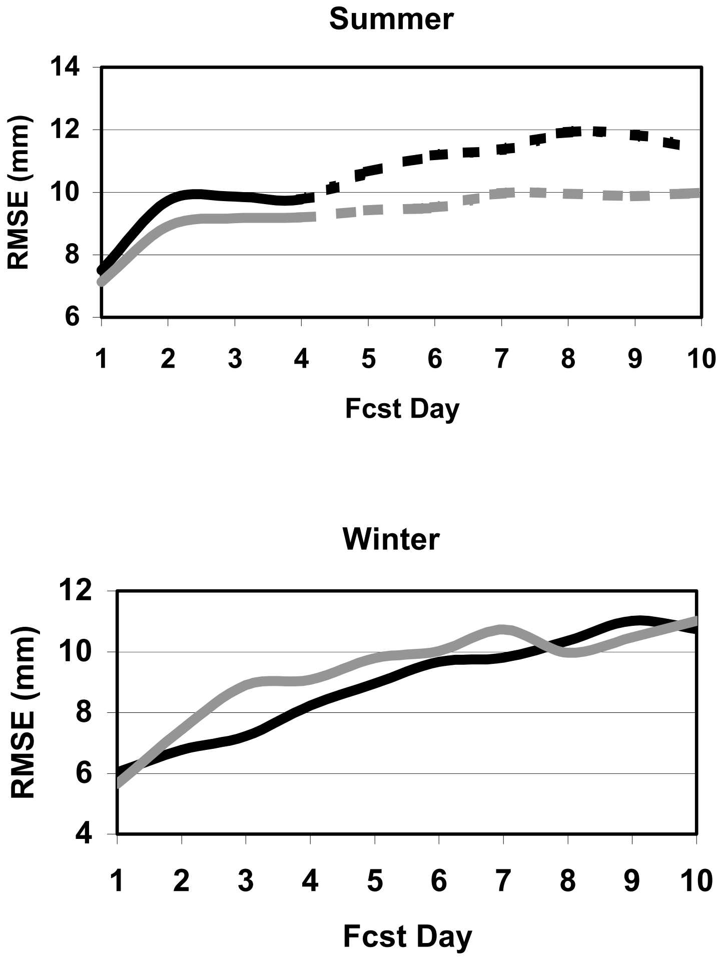


Fig. 21: Evolution of the root-mean-square error for precipitation amounts from the ECMWF high-resolution (TL319L31) deterministic forecasts. Top panel: the 1997 (gray) and 1998 (black) summers. Bottom panel: the 1997-98 (gray) and 1998-99 (black) winters.

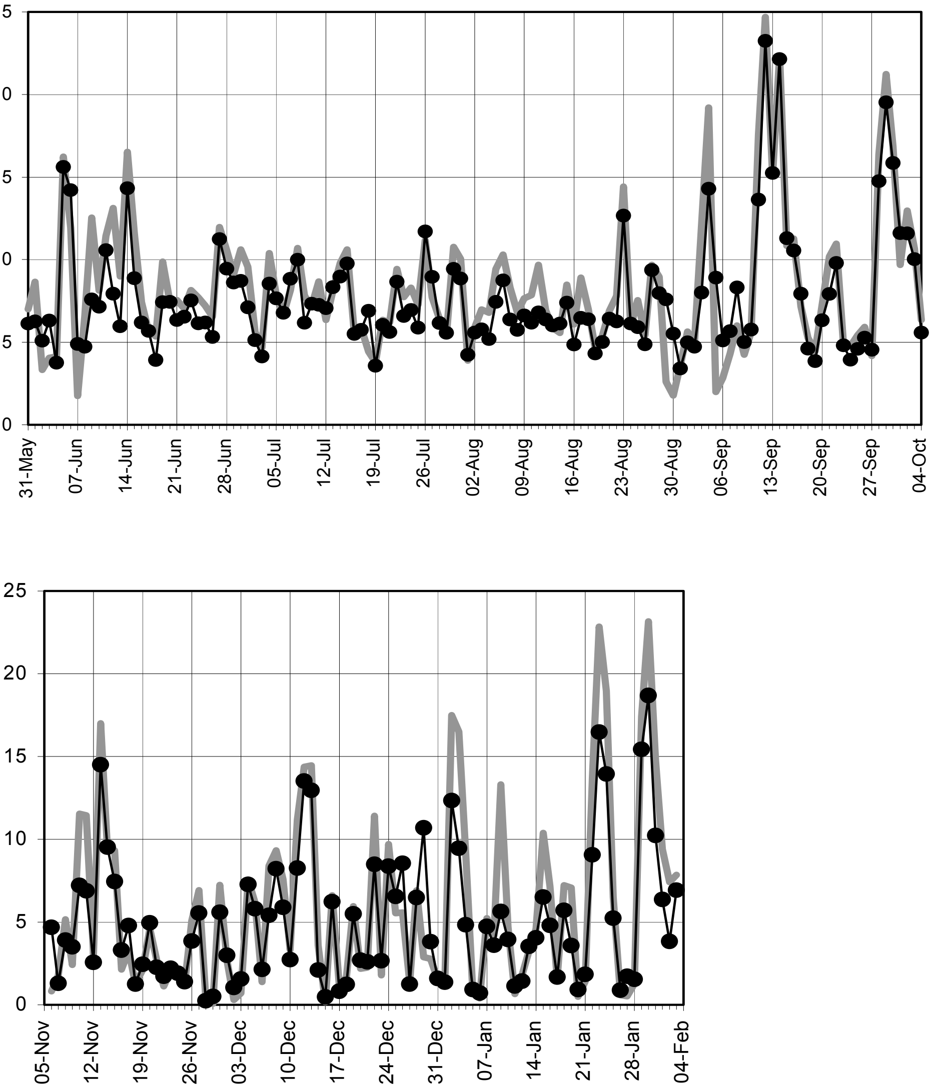


Fig. 22: Time series of daily RMSE for ensemble mean forecasts of precipitation, averaged for day 4 and day 5 forecasts valid at indicated dates (black line with dots), and spatial RMS for RFC verifying analyses (gray line). Top panel: 1998 warm season. Bottom panel: 1998-99 cool season.

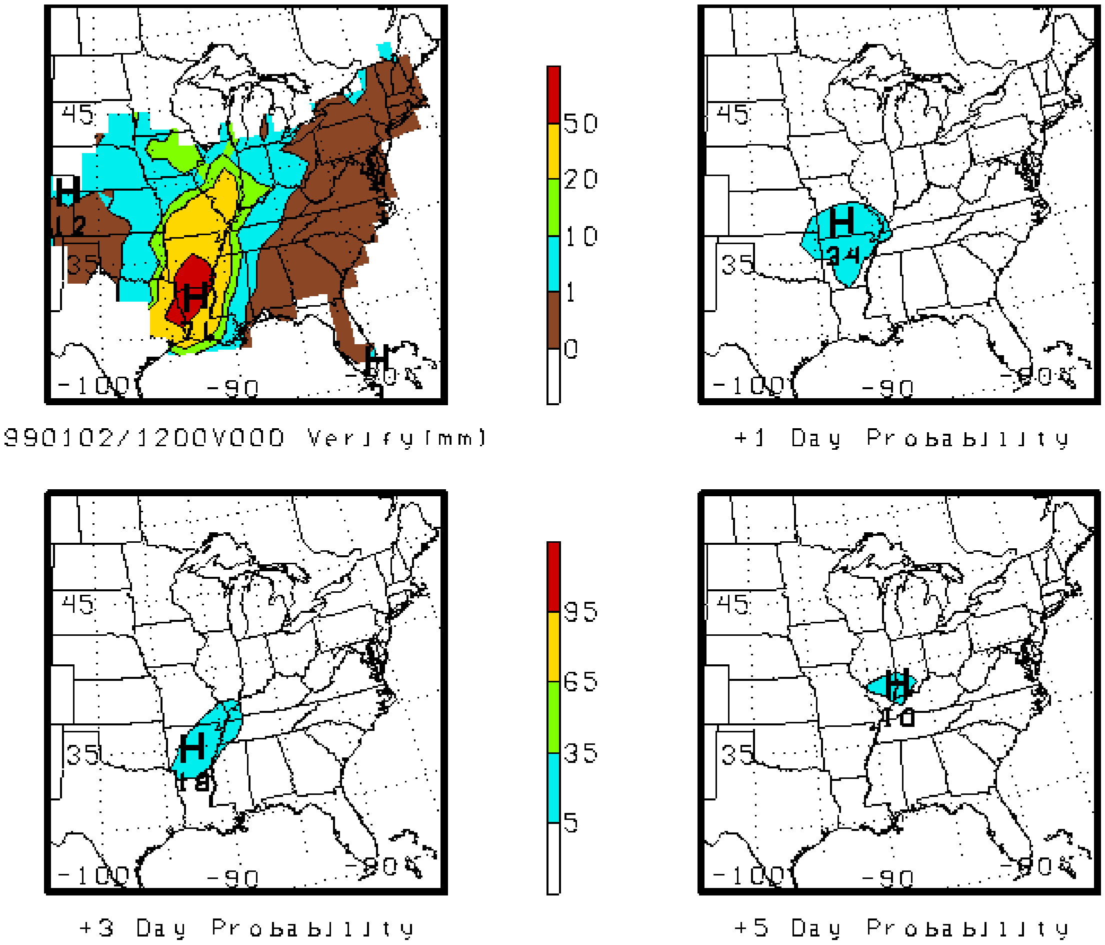


Fig. 23: 24-h accumulated rainfall, valid at 12 UTC 02 January 1999, for (a) verifying analysis, upper-left panel; and EPS probabilities of 24 h rainfall > 50 mm for (b) 1 day forecast, upper-right panel; (c) 3 day forecast, lower-left panel; and (d) 5 day forecast, lower-right panel. Shading in panel (a) for rainfall values of 0-1 mm, 1-10 mm, 10-20 mm, 20-50 mm and > 50 mm; no shading indicates grids with too few RFC stations to include in verification; legend at top portion of figure. Shading in panels (b)-(d) for probabilities of 1%, 25%, 50%, 75% and 99%; legend at bottom portion of figure.

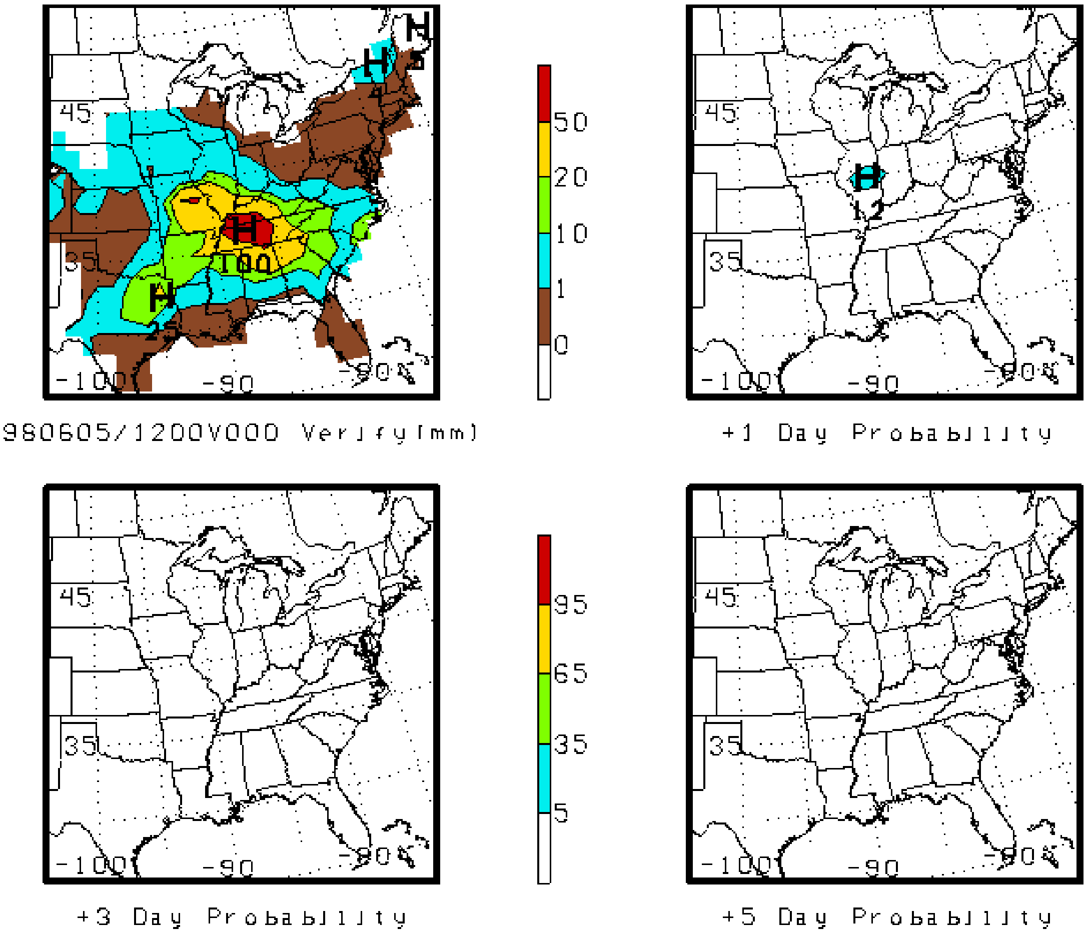


Fig. 24: As in Fig. 23 except for 5 June 1998.

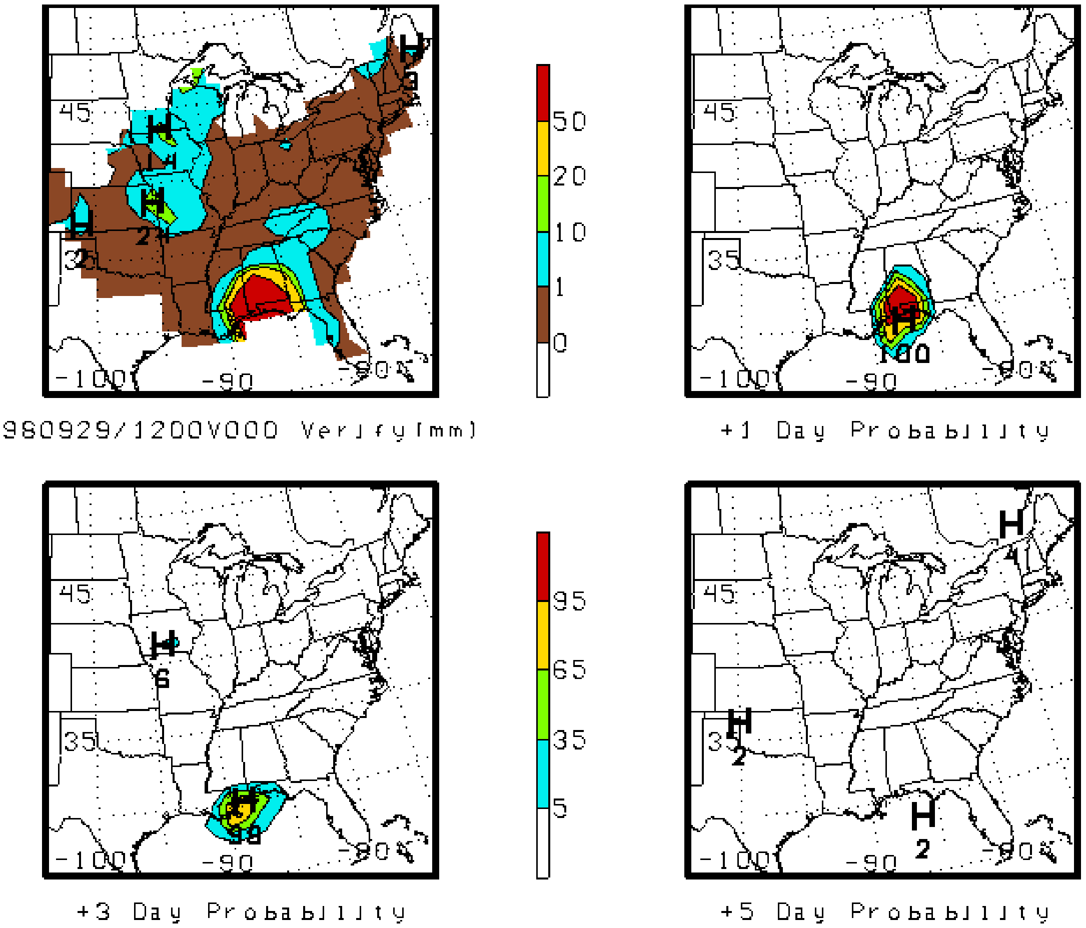


Fig. 25: As in Fig. 23 except for 29 September 1998, landfall of hurricane Georges.

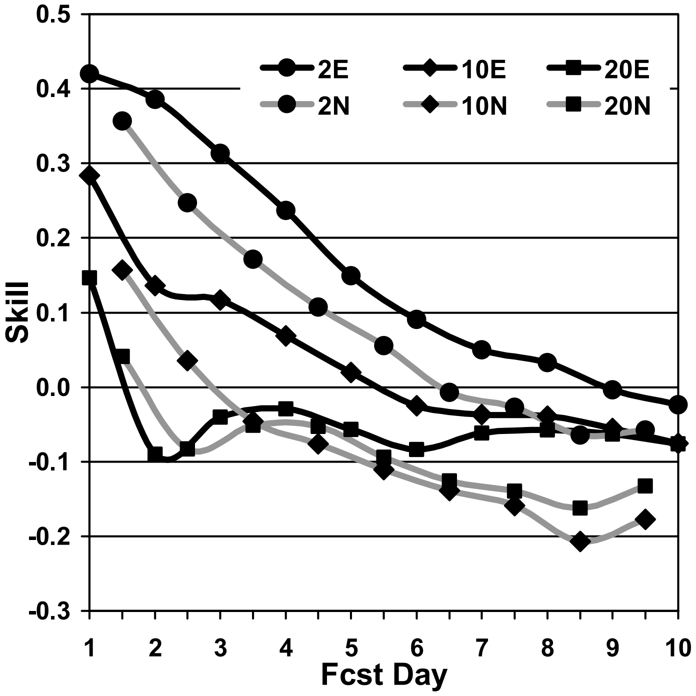


Fig. 26: Brier Skill Score for ECMWF EPS (lines with circles) and NCEP ensembles (lines with diamonds) for thresholds of 2 mm (black lines, black inserts), 10 mm (black line, gray inserts) and 20 mm (gray lines, black inserts). NCEP results taken from Fig. 11 of Eckel and Walters (1998); NCEP ensemble consists of 11 members, verified on a 2.5° by 2.5° degree grid, started from 0000 UTC analyses. ECMWF EPS results are based on only 11 ensemble members, spatially filtered to a 2.5° by 2.5° degree grid, but started from 1200 UTC analyses. See text for details.