A FORECAST RELIABILITY INDEX FROM ENSEMBLES: A COMPARISON OF METHODS

M.S. Roulston[1,2], C. Ziehmann and L.A. Smith[1,2]

[1] Pembroke College, Oxford University, Oxford, OX1 1DW, U.K.
[2] Centre for the Analysis of Time Series, London School of Economics, U.K.

Produced for Deutscher Wetterdienst

ABSTRACT

Two different approaches for constructing skill predictors for a deterministic forecast are compared. The first method uses a single statistic of the ensemble distribution to classify the deterministic forecast, the second method is to generate a completely probabilistic forecast from which both a deterministic forecast and a skill predictor can be extracted. The first method is evaluated using precipitation forecasts at five stations in Germany, the second method is evaluated using precipitation forecasts at the same five stations, and also temperature forecasts at 26 German stations. Both methods are found to produce useful *a priori* predictors of forecast skill. The second method, however, appears to have greater resolution without sacrificing reliability.

Contents

EXECUTIVE SUMMARY

It is now widely accepted that some *a priori* indication of forecast skill to accompany a forecast is desirable. This study evaluates two general approaches to obtaining skill predictions for a given forecast.

1. Obtain a relationship between forecast reliability and some statistic of an ensemble forecast.
2. Attempt to construct a probabilistic forecast from which a summary of forecast skill can be extracted.

The potential of both approaches was investigated for several locations in Germany using both the ECMWF ensemble prediction system and DWD medium range forecasts. It was found that both methods are feasible. The second method, however, offers several advantages:

a A more comprehensive probabilistic forecast contains more information than a single skill prediction. This information can be utilized by more sophisticated users.

b Probabilistic forecasts allow for greater flexibility in designing summary skill forecasts.

c It is possible to make more confident predictions of forecast skill from probabilistic forecasts.

The study found that probability forecasts of temperature based on the ECMWF ensemble prediction system have skill out to the full 10 day leadtime for which they are issued. This skill is reflected in both probabilitistic forecasts with a greater information content than a climatological forecast, and also in the ability to construct skill predictors which can identify bad forecasts *a priori.* It was also found that probabilistic forecasts constructed using only 6 members of the ECMWF ensembles typically contain almost as much information as the those generated from the complete, 51-member ensembles *when temperature at a single location is the variable under consideration.*

Treating the ECMWF ensemble distribution directly as a probabilistic forecast leads to *very poor* forecasts. These forecasts are actually worse than climatology. The ECMWF ensemble, however, does contain information about the state dependent predictability which can extracted and used to construct probabilistic forecasts using the methods described in this report.

Using both the ECMWF and DWD forecasts to construct probabilistic forecasts for precipitation does not lead to a significant improvement over using only the ECMWF ensemble prediction system.

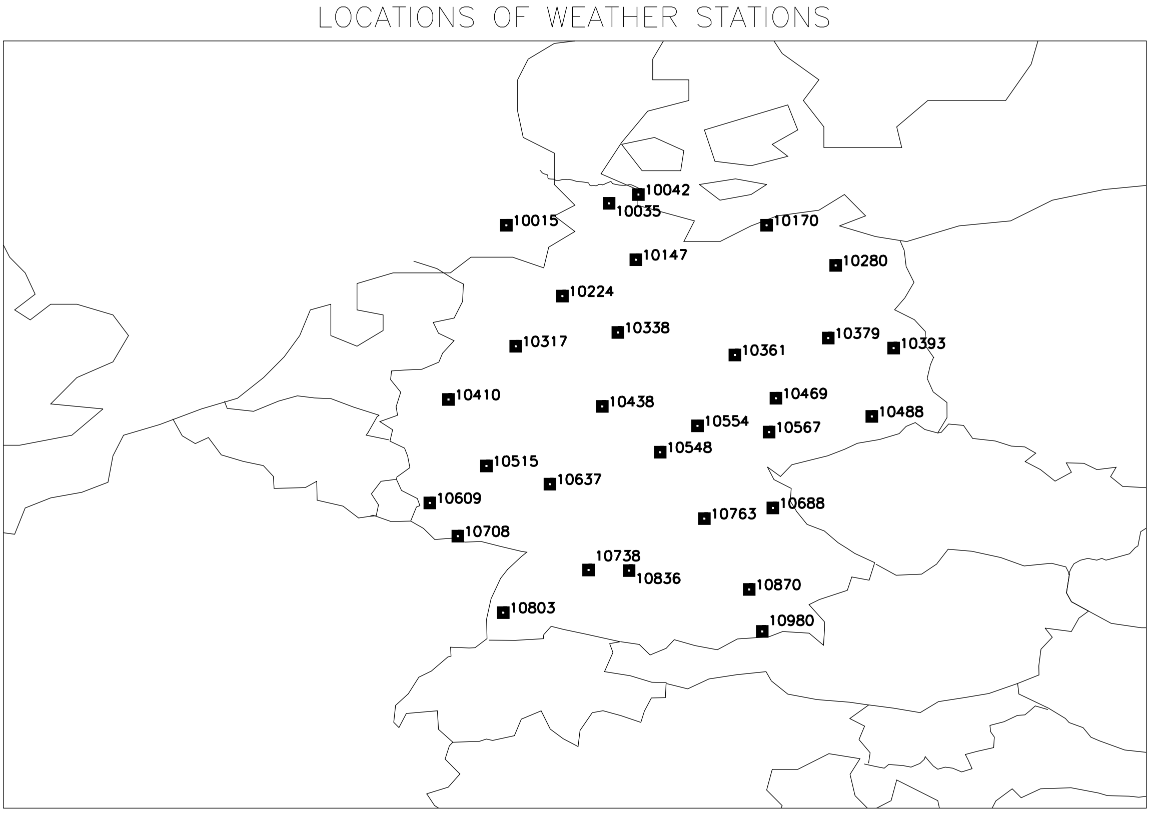


Figure 1: A map showing the locations of the weather stations used in this study.

Introduction

Traditional weather forecasts have consisted of single, *deterministic* forecasts. More recently, the potential value of an *a priori* estimate of forecast quality has been appreciated. Ensemble prediction systems offer an indicator of the flow dependent predictability of the atmosphere. Previous studies have investigated the ability of ensemble forecasts to provide estimates of forecast uncertainty [Buizza 1997; Toth et al. 2001; Ziehmann 2000; Ziehmann 2001]. The first part of this report is a continuation of the investigation by Ziehmann [Ziehmann 2000]. It investigates the use of ECMWF ensemble mode population, and the entropy of the ensemble distribution, as predictors of forecast skill for precipitation forecasts at five stations in Germany. The locations of all the stations used in the study are shown in Fig. 1.

The second part of this report presents two different methods for constructing probilistic forecasts from the information contained in the dynamical forecasts—both the high resolution forecasts issued by ECMWF and DWD and also the ensemble forecasts produced by ECMWF. The methods are applied to temperature and precipitation forecasts. Finally, the probabilistic forecasts are used to constructs skill predictors for deterministic forecasts, and the reliability of these skill predictors are evaluated.

1 Skill predictors for precipitation

1.1 Method

In Ziehmann (2000) the population of the mode of the ensemble distribution, along the entropy of this distribution, were identified as being good predictors of the probability that a given temperature forecast would be correct. This method will now be extended to precipitation forecasts for five station in Germany. For the case of temperature, climatologically equi-probable bins were used to quantize the forecasts. This cannot be done with precipitation due to the high fraction of days on which there is no precipitation. Three precipitation (ppt) categories were defined; $[\text{ppt} = 0, 0 < \text{ppt} \leq 3\text{mm}, \text{ppt} > 3\text{mm}]$. Success was defined when the ECMWF high resolution forecast fell in the same bin as the observation. To provide a baseline for the success rate, success rates for 50 shuffled sequences of the forecasts were determined. These provided a distribution for the *baseline success rate.* This baseline is necessary because, due to the high frequency of zero precipitation days, a reasonably high success rate can be achieved by issuing random forecasts drawn from climatology. For each forecast the ensemble forecast was binned. Let f_i be the fraction of ensemble members in the i^{th} bin (where $i = 1, 2, 3$). The mode population and the negentropy [1] were calculated. The ensemble mode population is proportional to $\max_i f_i$, while the negentropy of the ensemble distribution is given by $\sum_i f_i \log_2 f_i$. Ensemble mode population and negentropy are both potential predictors of forecast skill. Henceforth, the word "predictor" will be used to refer to them both. A low threshold was chosen to be at the r^{th} percentile of the distribution of the predictor, while a high threshold was chosen to be at the $(100 - r)^{\text{th}}$ percentile. Forecasts for which the predictor fell below the low threshold were classed as having low predictability, while those for which the predictor exceeded the high threshold were classed as having high predictability. The success rates for the low and high predictability forecasts were calculated separately. To estimate uncertainty in these success rates bootstrap resampling was used. Fifty sets of forecasts were constructed by resampling, with replacement, from the original sample. The success rate associated with each set of forecasts was calculated giving an estimate for the mean success rate, along with an estimate for the standard deviation in this estimate [Efron and Tibhsirani 1986].

1.2 Results

Figures 2 and 3 show the results for five German stations obtained when using mode population and negentropy respectively as predictors of forecast skill. All the panels show results for 5-day precipitation forecasts. The success rate for both high and low predictability cases at the 100^{th} percentile is the average success rate for the forecasts. It is higher than the baseline success rate for all stations. In all cases if the predictor exceeds its 90^{th} percentile the success rate is approximately double the baseline success rate, whereas if the predictor falls below its 10^{th} percentile the success rate falls below the baseline success rate. Except for station 10469, where negentropy provides a slightly better separation of success rates than mode population, there is little difference between these two predictors of forecast skill.

[1] Negentropy (negative of the entropy) was used solely so that, in both cases, a high value of the predictor is an indicator of high predictability.

At the stations examined it typically does not rain on 40% of days. This is reflected in relatively high base line success rate of the precipitation forecasts. The precipitation forecasts were separated into days for which rain was forecast and days for which no rain was forecast. The ability of the ensemble distribution mode population to predict forecast skill was then evaluated separately for these two cases. For this part of the analysis a simpler binary binning, $[\mathrm{ppt} = 0, \mathrm{ppt} > 0]$. The results are shown in Fig. 5 (5-day rain forecasts), Fig. 6 (3-day rain forecasts) and Fig. 7 (5-day no rain forecasts). Note that in Figs. 5-7 the base line success rate is equal to the average success rate since all forecasts are identical. Furthermore, since reordering identical forecasts does not change the success rate there the base line has no thickness.

Figure 5 indicates that even if a forecast of rain is classed as "high predictable", using a restrictive definition (low percentile), the mean probability of it raining very rarely exceeds about 70%. This means, that if this method is used to classify *a priori* predictability, it will not be possible to issue a forecast of rain with a probability of more than about 70%. Figure 6 shows the same result, but this time for the 3 day forecast. For stations #10379, #10410 and #10637 using the most restrictive definition of high predictability enables forecasts to be made with confidences of almost 90%. However, at the other two stations the success rate of the predictable forecasts is still 70% or less. Figure 7 shows that this relatively poor ability to be able to forecast rain with confidence is compensated for by the ability to make highly confident forecasts of no rain. Indeed, if a forecast of no rain is accompanied by a mode population in the top 10^{th} percentile it is almost certain not to rain. A similar result has been previously obtained for temperature forecasts, with high predictability forecasts having a success rate of around 70% [Ziehmann 2000; Ziehmann 2001].

Figure 4 shows the results of using the ECMWF ensemble mode population as a predictor of the skill of the kalman filtered DWD forecast.

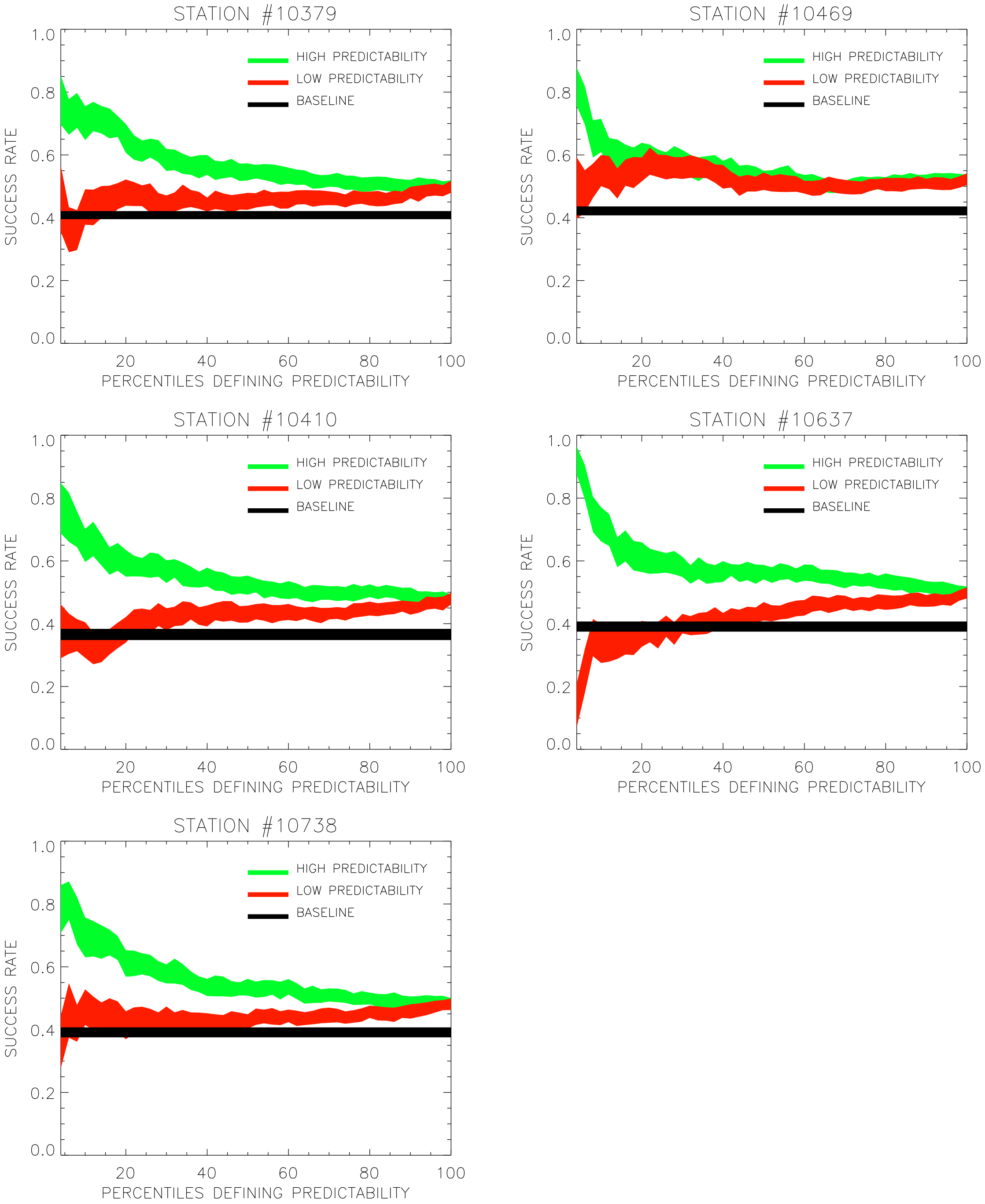


Figure 2: Performance of high and low predictability 5-day precipitation forecasts. High (low) predictability was defined as having an ensemble mode population above (below) the chosen percentile. The black line is the baseline success rate. The thicknesses of all the lines are two standard deviations obtained using bootstrap resampling.

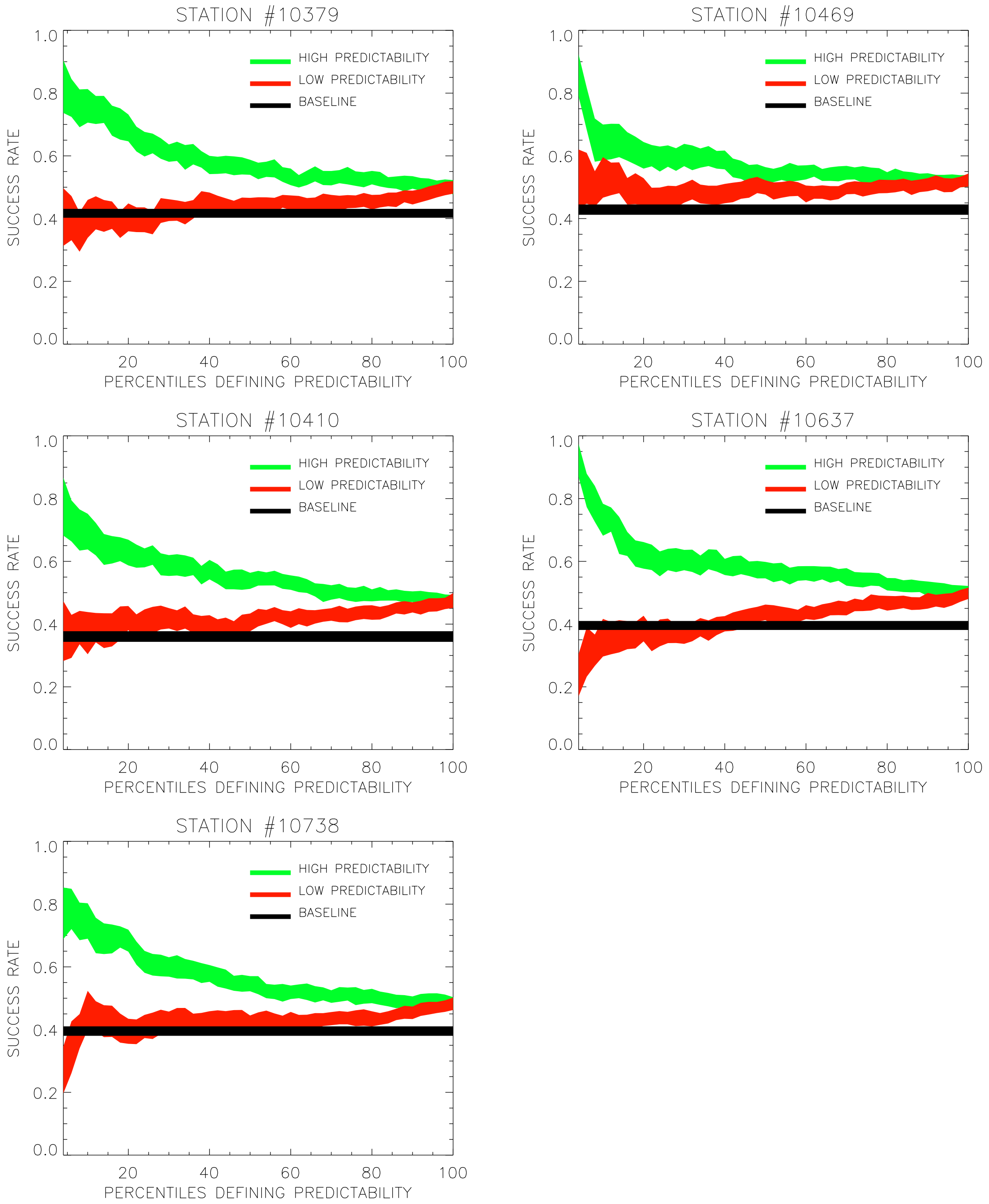


Figure 3: As Fig. 2 but using negentropy of the ensemble distribution as the predictor of predictability rather than mode population.

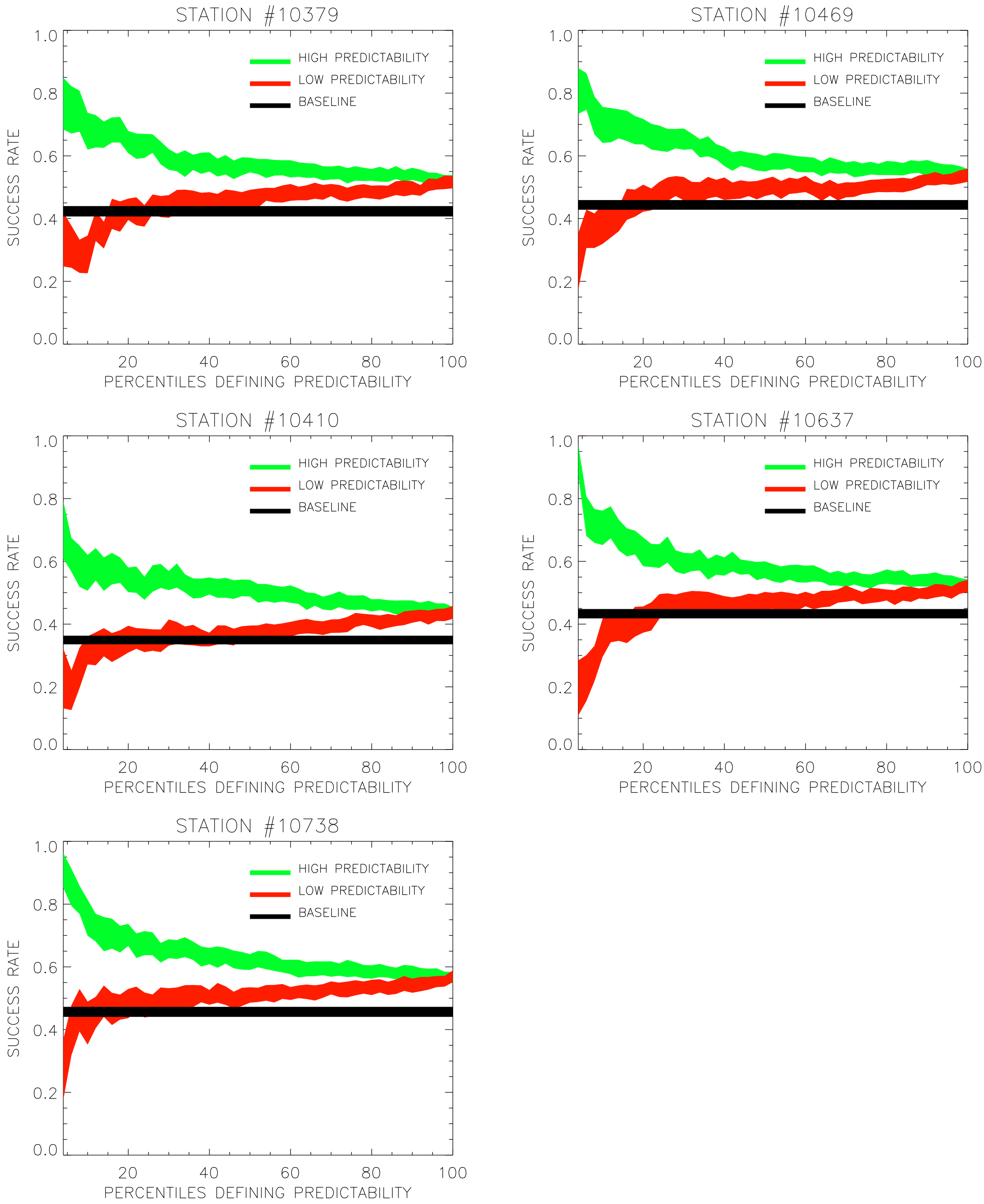


Figure 4: As Fig. 2 but using the kalman filtered DWD forecast instead of the high resolution ECMWF forecast—the ECMWF EPS was still used to obtain the skill predictor.

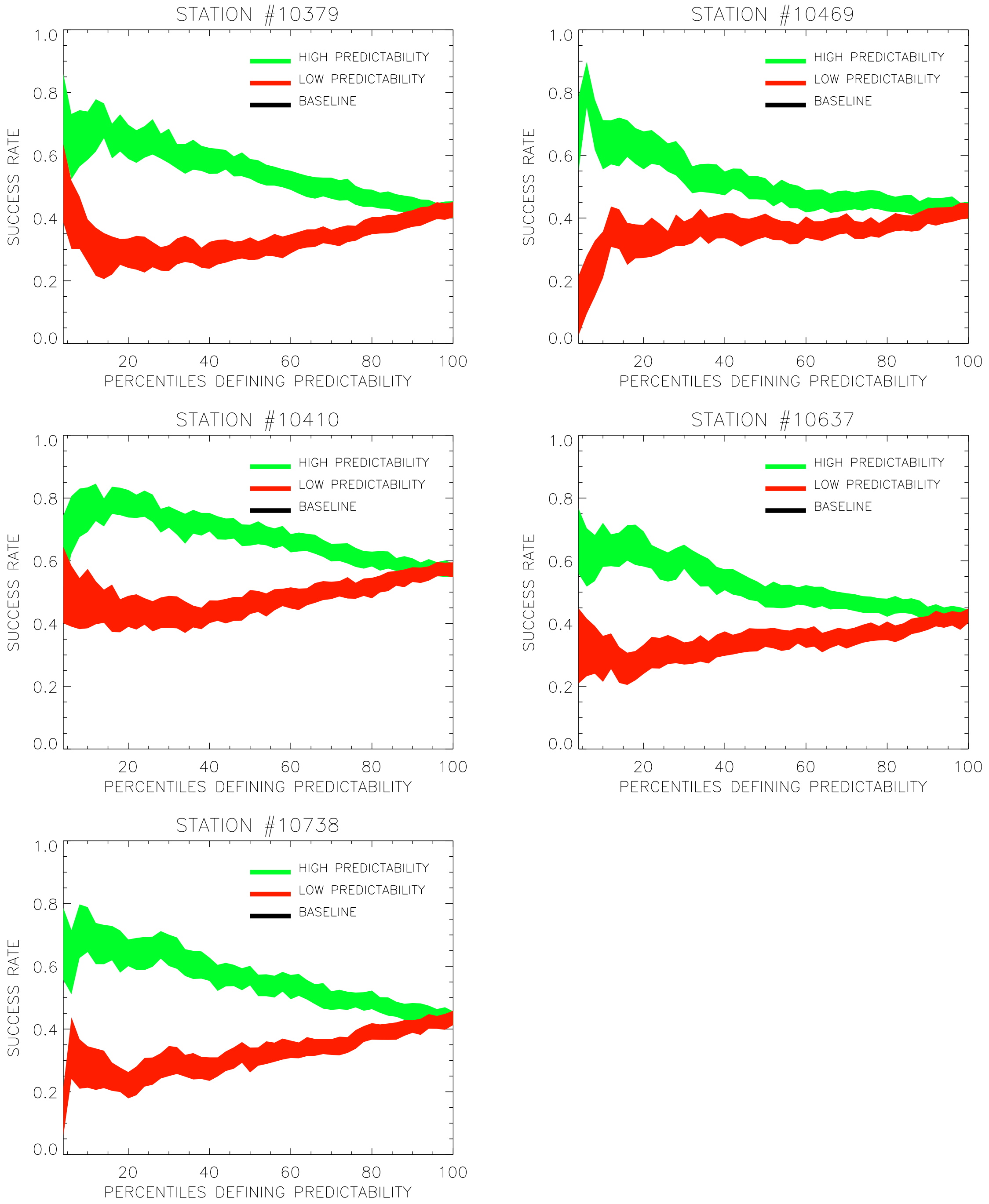


Figure 5: As Fig. 2 but only using days for which rain was forecast, and only using a binary binning of rain/no rain.

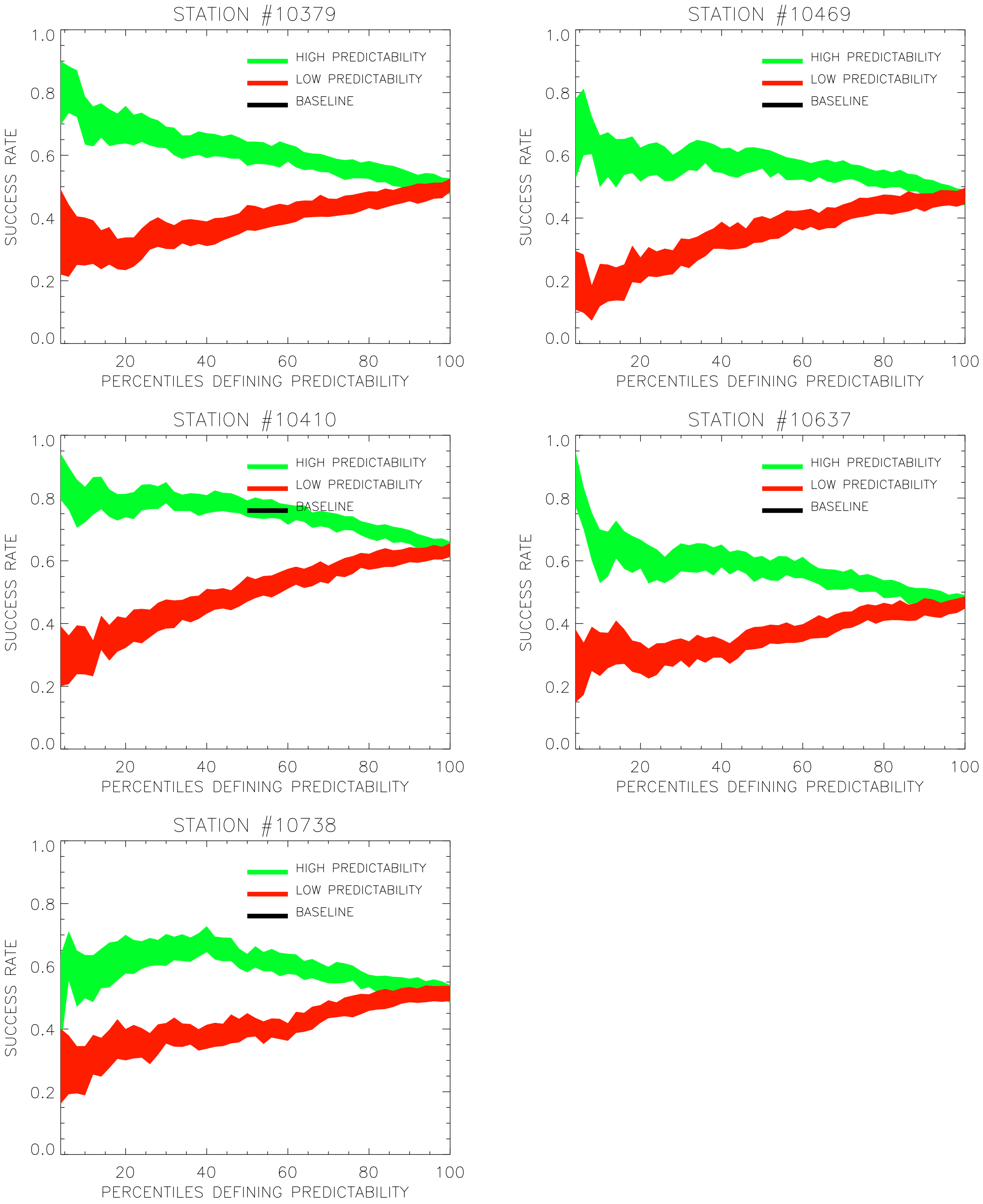


Figure 6: As Fig. 2 but for the 3 day forecast, and only using days for which rain was forecast, and only using a binary binning of rain/no rain.

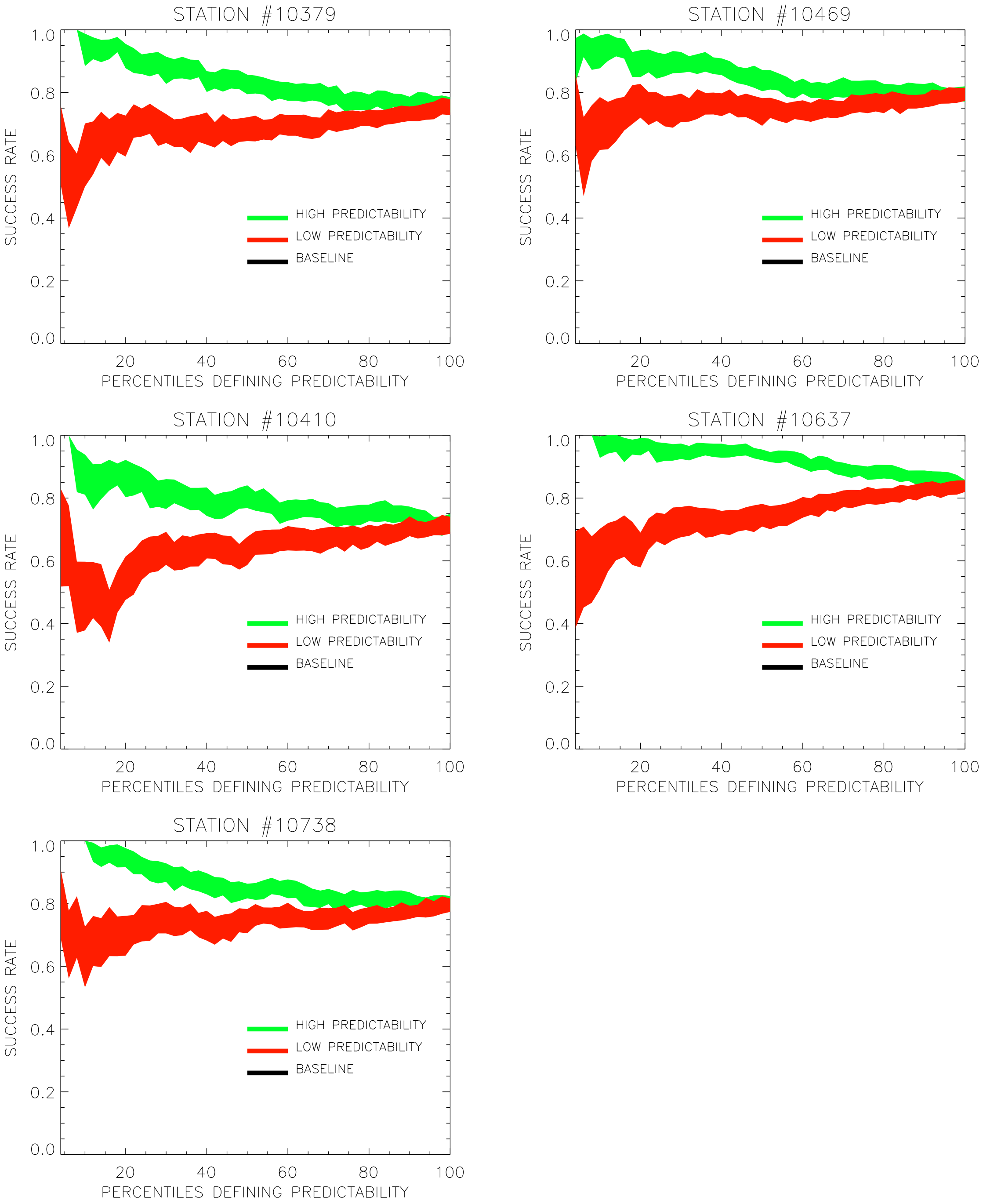


Figure 7: As Fig. 2 but only using days for which no rain was forecast, and only using a binary binning of rain/no rain.

2 Probabilistic forecasting

2.1 Motivation

Probabilistic forecasts are the most complete type of forecast. Such a forecast encompasses all known information about forecast uncertainty. This is the ideal forecast product for sophisticated users who can quantify the weather dependence of their utility and their risk tolerance. If a forecast centre issues only a deterministic forecast, even one accompanied by a prediction of skill, they are making an *implicit* assumption about users' utility functions and risk tolerances—an assumption that will almost certainly be wrong [Smith et al. 2001].

In this section two methods for constructing probabilistic forecasts from dynamical forecast products are presented and evaluated. The first method is *best member dressing*, in which each member of an ensemble forecast is convolved with the distribution of the historical "best member" errors. The second method is the *method of analogues*, in which historical analogues of the current forecast are found. The observations corresponding to these analogues are treated as a sample of the forecast probability distribution.

2.2 Evaluating probabilistic forecasts

The evaluation of probabilistic forecasts is a relatively underdeveloped field. In this study two *scoring rules* were used to assess the quality of the probabilistic forecasts. The first was a quadratic rule, commonly called the Brier score [Brier 1950]. For m possible outcomes the Brier score is defined as

$$BS = \sum_{i=1}^{m} (p_i - \delta_{ij})^2 \tag{1}$$

where p_i is the forecast probability of outcome i, the actual outcome is j and $\delta_{ij} = 0$ if $i \neq j$ and $\delta_{ij} = 1$ if $i = j$. The second scoring rule was a logarithmic rule which corresponds to the information deficit or *ignorance* of someone in possession of the forecasts [Roulston and Smith 2001a]. This is defined as

$$IGN = -\log_2 p_j \tag{2}$$

where again, j is the actual outcome, which had a forecast probability, p_j. Both the quadratic Brier score and the logarithmic ignorance score are *proper* scoring rules—the forecaster achieves the best expected score by reporting their true estimates of the probabilities.

Another method for scoring probabilistic forecasts is the cost-loss method [Katz and Murphy 1987; Katz and Murphy 1997; Richardson 2000]. This method uses the realized loss of a loss minimizing user, with a given cost-loss (C/L) ratio, as a skill score. It can be shown that the Brier score corresponds to this realized utility averaged over a distribution users which is uniform in C/L [Murphy 1966; Richardson 2001]. It can also be shown that the ignorance corresponds to the realized loss averaged over a distribution of users given by [Roulston and Smith 2001a]

$$u(C/L) \propto \frac{1}{(1 - C/L)(C/L)} \tag{3}$$

The derivation of the relationships between cost-loss and Brier score and ignorance can be found in the Appendix. Neither Brier score nor ignorance is the "best" scoring rule in any general sense. In some ways, they are complementary, with ignorance being more sensitive to extreme probabilities, and Brier score being relatively more sensitive to moderate probabilities.

2.3 The best member dressing method

2.3.1 Description of the method

The best member dressing method was used to construct probabilistic temperature forecasts from ECMWF ensembles. To make direct comparisons between the single, high resolution forecast generated by ECMWF and the ensemble prediction system, the single forecasts were converted into ensembles. This was done using historical errors for the base period May 1997 to April 1998. Each high resolution forecast for the test period (May 1998 to April 1999) was converted to a 102 member ensemble by adding 102 historical forecast errors, picked randomly with replacement, from the 102 days in the base period which had an annual phase closest to the forecast under test. These ensembles are of the type commonly referred to as *statistical ensembles.* Statistical ensembles can be compared with *dynamical ensembles.* However, it has been shown that better results can be obtained if the dynamical ensembles are used to construct hybrid dynamical-statistical ensembles or *hyperensembles.* In this study the hyperensembles were constructed by dressing each member of the dynamical ensembles with the "best member" errors as previously described by Roulston and Smith [2001b, 2001c].

Let $\mathbf{y}$ be an m-component vector describing "truth". Truth can be decomposed into a *dynamical* component, $\mathbf{x}$, and a *statistical* component, ε.

$$\mathbf{y} = \mathbf{x} + \varepsilon \tag{4}$$

Note that the terms *dynamical* and *statistical* are being used in an operational sense. The dynamical component contains all processes contained in the forecasting model. The statistical component includes all contributions that are to be dealt with statistically, such as model error and residual initial condition error. The dynamical component need not be strictly deterministic if, for example, the model contains stochastic parameterizations. Conversely, the statistical component will include processes that are really deterministic, but which will be dealt with in a statistical manner. Let $\mathbf{x}_i$ be an ensemble of N dynamical forecasts $(i = 1, \ldots, N)$. The best member of the ensemble is the member that has the "correct" dynamical component. If truth is known, and the best member is identified, the contribution of ε can be determined. Over many forecasts, the statistical properties of the ε component can be estimated. For the following analysis, it will be assumed that ε has a multivariate normal distribution with uncorrelated components. The probability that the i^{th} ensemble member has the correct dynamical component is

$$p_i = \frac{\exp\left(-\sum_{k=1}^{m}(y_k - x_{i,k})^2/2\sigma_k^2\right)}{\sum_{j=1}^{N}\exp\left(-\sum_{k=1}^{m}(y_k - x_{j,k})^2/2\sigma_k^2\right)} \tag{5}$$

where y_k is the k^{th} component of $\mathbf{y}$, and $x_{i,k}$ is the k^{th} component of $\mathbf{x}_i$. If $\mathbf{x}_l$ is the correct dynamical component then $E[(y_k - x_{l,k})^2] = \sigma_k^2$. If Ω_k^2 is the variance of the k^{th} component of all the ensemble members then an approximate expression for p_l, the probability assigned to the true best member, is

$$p_l \sim \frac{\exp(-m/2)}{\exp(-m/2) + (N-1)\exp\left(-\sum_{k=1}^{m}(\Omega_k^2 + \sigma_k^2)/2\sigma_k^2\right)} \tag{6}$$

To simplify Eq. 6 assume that all the σ_k are identical and all the Ω_k are also the same.

$$p_l \sim \frac{\exp(-m/2)}{\exp(-m/2) + (N-1)\exp(-m/2 - m\Omega^2/2\sigma^2)} \tag{7}$$

If $N \gg 1$ Eq. 6 becomes

$$p_l \sim \frac{1}{1 + N\exp(-m\Omega^2/2\sigma^2)} \tag{8}$$

An examination of Eq. 8 provides some insight into the conditions required for the correct identification of the best member. For an unambiguous identification $p_l \approx 1$. Therefore, it is required that

$$N\exp(-m\Omega^2/2\sigma^2) \ll 1 \tag{9}$$

From Eq. 9 it can be seen that if $\Omega \gg \sigma$ then the best member can be easily identified. That is, if the spread of the ensemble members is much greater than the uncertainty due to the statistical component, then the ensemble member that comes closest to truth is highly likely to be the best member. If, however, $\Omega \approx \sigma$ and m is small then this is not the case. This is because if the size of ε is comparable to the spread of the $\mathbf{x}_i$, then the $\mathbf{x}_i$ closest to $\mathbf{y}$ is not necessarily the correct $\mathbf{x}$. In this situation, the chance of choosing the correct best member can be improved by increasing m. That is, the best member is identified as the closest $\mathbf{x}_i$ to $\mathbf{y}$ in a higher dimensional space. In practice, this can be done by comparing the forecast to truth at multiple points in space and time. This means, that even if, one is only interested in forecasting a univariate quantity, the best member must be chosen on the basis of a multivariate forecast. From Eq. 9, it can also be seen that the probability of correctly identifying the best member falls as the ensemble size, N, increases. This makes sense; the more ensemble members there are, the higher the chance of misidentifying the best member. If N is increased then m should also be increased to ensure that the best member of the enlarged ensemble is correctly identified.

In practice, σ is a statistical property of ε, and so will not be known until ε has been estimated. The problem is to decide, *a priori*, whether enough variables are being used to accurately identify the best member of an ensemble. To do this the idea of a *false best member* is introduced. Let the normalized distance between the i^{th} ensemble member, $\mathbf{x}_i$, and truth, $\mathbf{y}$, in the space of d variables be written as $R_{i,d}$ where

$$R_{i,d}^2 = \sum_{k=1}^{d} \frac{(x_{i,k} - y_k)^2}{\Omega_k^2} \tag{10}$$

Note that in Eq. 10 the k^{th} component has been normalized by the ensemble standard deviation of that component, Ω_k. The best member, in this d-dimensional space, is the one which has the minimum $R_{i,d}^2$. If this best member is the true best member, then it should remain the best member when an extra variable is added. That is, if $\min R_{i,d}^2 = R_{j,d}^2$ then $\min R_{i,d+1}^2 = R_{j,d+1}^2$. If this condition doesn't hold then the best member can be classed as a *false* best member (FBM). The fraction of FBMs, averaged over past forecasts, gives an indication of whether d is high enough. The variables that are added can either be the same quantity at different spatial locations or they can be at the same location, but for different forecast lead times. Different forecast quantities can also be used.

After identifying the best member of an ensemble, the error of this ensemble member can be calculated. The covariance matrix of the multivariate errors can be calculated using a set of past forecast-truth pairs. This covariance matrix can be used to generate *hyperensembles.* This is done by dressing each ensemble member with a *daughter ensemble*, generated using the best member error statistics.

As described above the ensemble forecasts for temperature were generated using two methods:-

(i) The high resolution "best guess" forecast was dressed with a statistical ensemble constructed from by selecting 102 forecast errors picked randomly, with replacement, from the 102 days in the base period closest with an annual phase closest to the forecast day. This will be referred to as the HIRES-based forecast.

(ii) Each member of the ECMWF dynamical ensemble forecast was dressed with a daughter ensemble constructed by from best member errors picked randomly, with replacement, from the 102 days in the base period closest with an annual phase closest to the forecast day. This will be referred to as the EPS-based forecast.

As well as using the full ECMWF dynamical ensemble ($N = 51$), truncated dynamical ensembles with $N = 6$ and $N = 17$ were also tested. The size of the daughter ensembles used in (ii) was $102/N$, thus the statistical ensembles of (i) and the hyperensembles of (ii) all had 102 members.

The application of the formulae given by Eqs. 1 and 2 requires that the outcomes be discrete. The ensemble forecasts (statistical and hyperensembles) were discretized by binning into 2°C bins. With the logarithmic (ignorance) score there is an issue concerning the assignment of zero probabilities to a particular outcome since, should this outcome occur, an infinite score will result. "Cromwell's Rule" [2] warns against assigning a zero probability to an event unless it is truly impossible [Lindley 1985]. In this case, a fictitious 103rd ensemble member was spread evenly across all bins to approximate the uncertainty in the forecast probabilities due to the finite size of the ensembles.

The value of each type of skill score was calculated for each forecast and observation. The mean value of each skill score was then estimated using a bootstrap resampling method [Efron and Tibhsirani 1986]. The series of skill scores was split into blocks, each 10 days in length. The mean value for each block was calculated. Fifty estimates of the mean value of the entire series were then made by resampling, with replacement, from the block means. The mean and the standard deviation of these fifty estimates were then calculated.

2.3.2 Results for best member dressing

Figures 8-11 show the average Brier scores for the probabilistic forecasts constructed from the single high resolution ECMWF forecasts, and for those constructed using the ensemble prediction system. The number of ensemble members used to constructed the EPS-based forecasts were 51, 17, 6 and 2 respectively. Figures 12-15 show the average ignorance scores for the HIRES-based probabilistic forecasts, and those made using the EPS forecasts. The number of ensemble members used for the EPS-based forecasts were 51, 17, 6 and 2 respectively. Both the Brier and ignorance scores have been subtracted from the corresponding score obtained from an estimate of the climatological distribution. Thus, zero (the dashed line) represents the skill of climatology, and any score less than zero signifies greater skill than climatology. The thicknesses of the lines in Figs. 8-15 are two standard deviations in the estimate of mean, obtained using the bootstrap resampling method described above. In all the figures the black line is for the forecasts made using the ECMWF high resolution forecast and the

[2] "I beseech you, in the bowels of Christ, think it possible you may be mistaken." [Oliver Cromwell in a Letter to the General Assembly of the Church of Scotland, 3 Aug 1650]

grey line is for the EPS-based forecasts. From the Figs. 8-15 it can be seen that the 51 member EPS-based forecasts provide significantly better probabilistic forecasts than the HIRES-based forecasts and almost every station evaluated. This result holds whether the quadratic Brier score, or the logarithmic ignorance score, are used for the evaluation. Furthermore, it can be seen that, even is only 17 members of the ECMWF dynamical ensembles are used to construct the EPS-based forecasts, the EPS-based forecast is still significantly better than the HIRES-based forecast. The EPS-based forecast still has a statistically significant edge over the HIRES-based forecast when only 6 members are used. When the EPS-based forecast is generated using only two dynamical ensemble members, however, there is little difference between the quality of the EPS-based forecasts and the HIRES-based forecasts. An inspection of Figs. 8 and 12 indicates that, while the skill of the HIRES-based forecasts falls to climatological skill at a lead time of about 180 hours, the EPS-based forecasts still have skill beyond climatology at 240 days. This result implies that EPS-based forecasts would still be more skillful than climatology, even if the forecast lead time were to be extended beyond 10 days. In this context, "skill" means that the probability distribution from the EPS-based forecast is better than a climatological distribution. For example, consider station #10015 in Fig. 12. At 240 hours the ignorance of the EPS-based forecast is about 0.2 bits less than that of the climatological forecast. This means that, if fair odds were set using the climatological probabilities (or the HIRES-based forecast probabilities) then, someone betting their wealth in proportion to the probabilities of the EPS-based forecast would, on average, increase their wealth by a factor of $2^{0.2} = 1.15$ per bet placed.

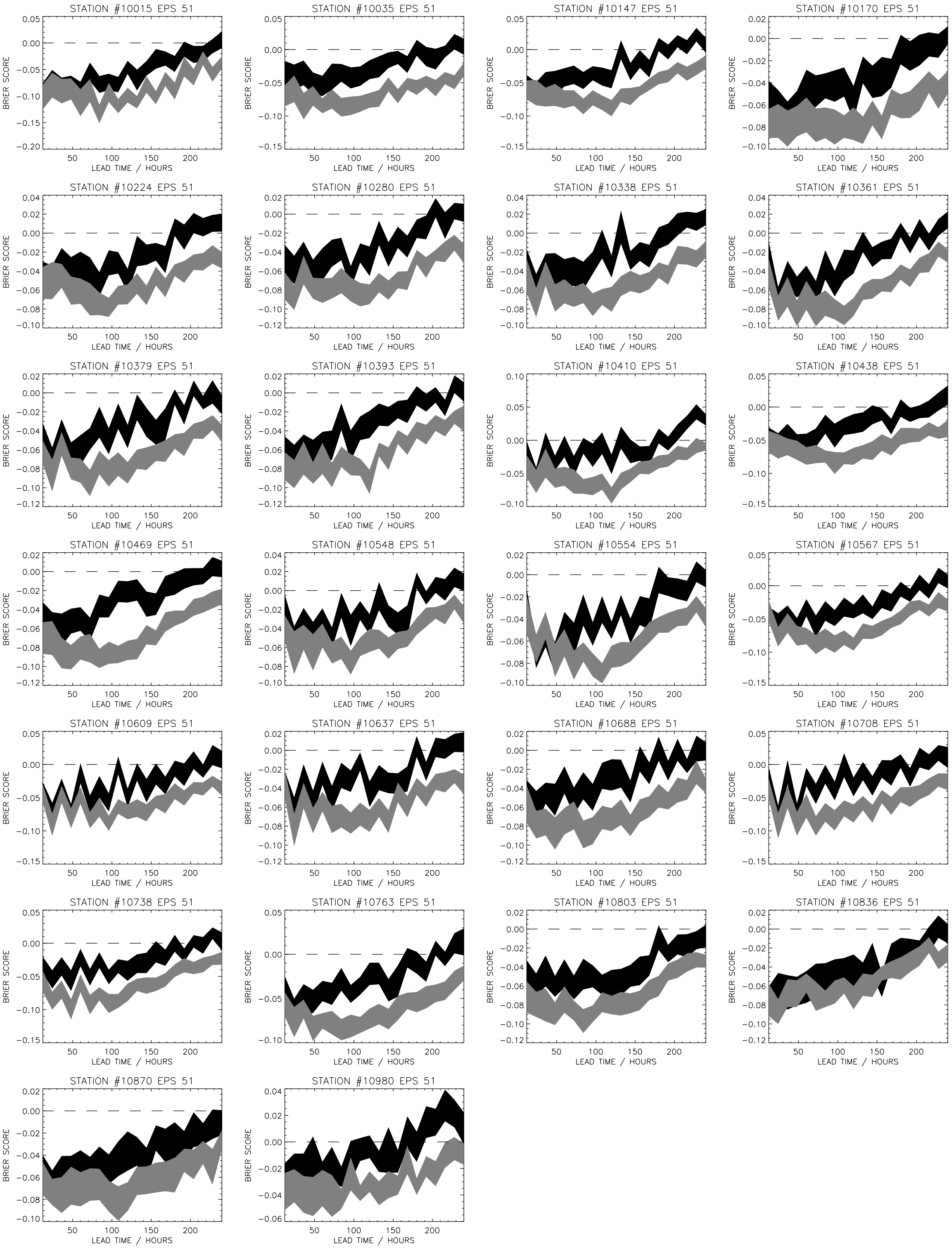


Figure 8: Average difference between quadratic (Brier) skill scores of the temperature forecasts and the skill score of a climatological distribution. Best guess forecasts with a statistical ensemble are in black and the hyperensembles constructed using the full 51 member ECMWF ensemble are in grey. The dashed line represents the skill of climatology.

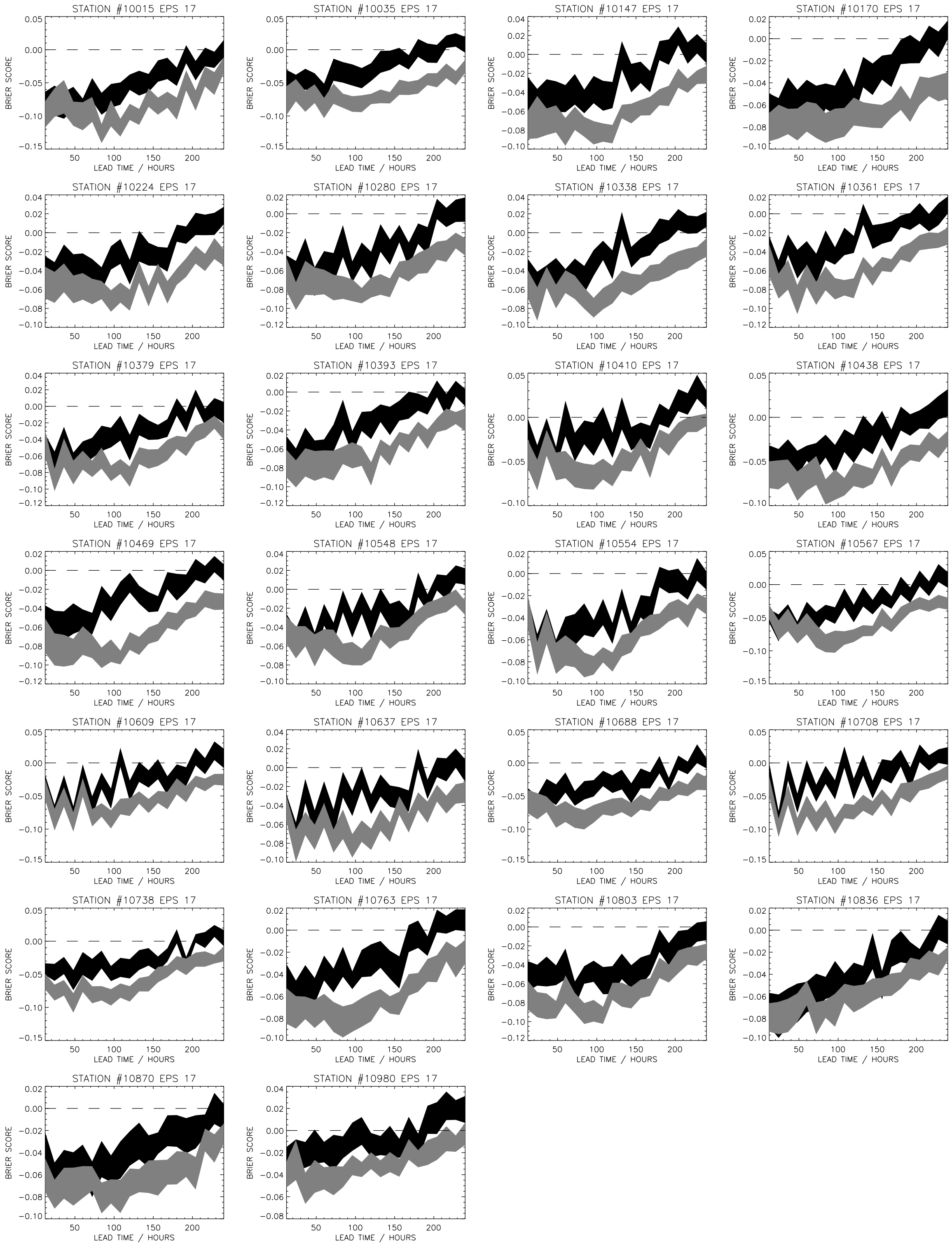


Figure 9: As Fig. 8 but the hyperensembles were constructed using only 17 members of the ECMWF dynamical ensemble.

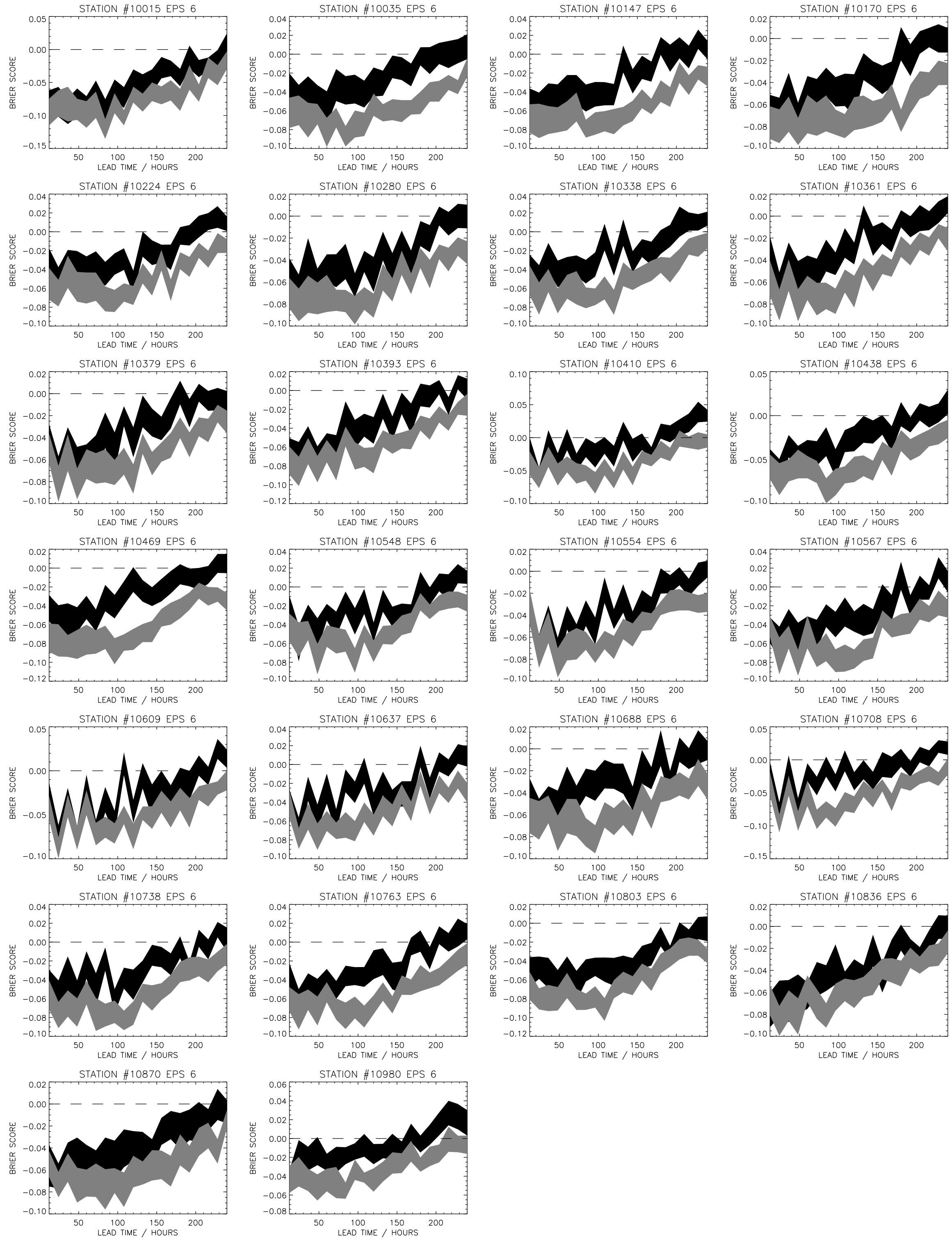


Figure 10: As Fig. 8 but the hyperensembles were constructed using only 6 members of the ECMWF dynamical ensemble.

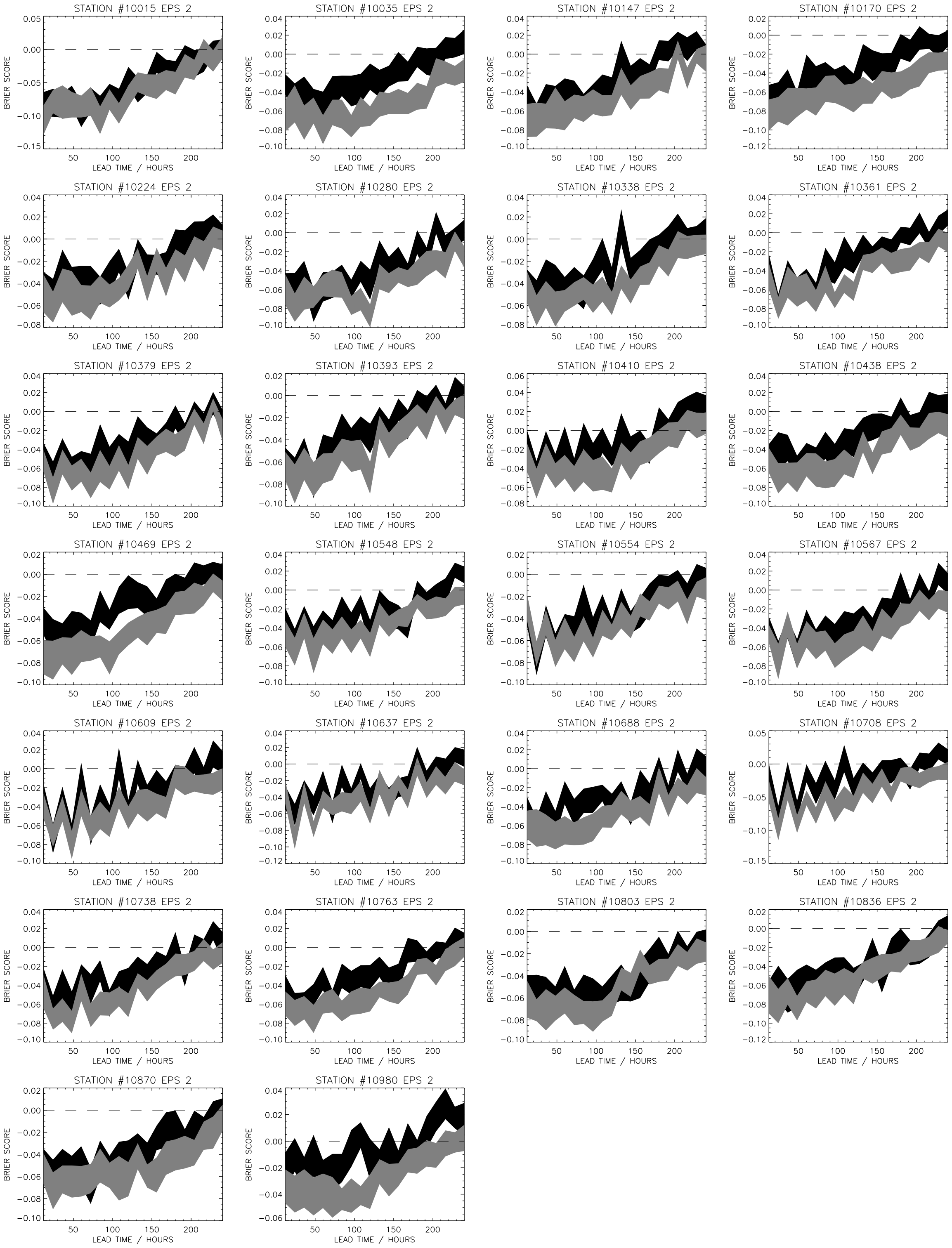


Figure 11: As Fig. 8 but the hyperensembles were constructed using only 2 members of the ECMWF dynamical ensemble.

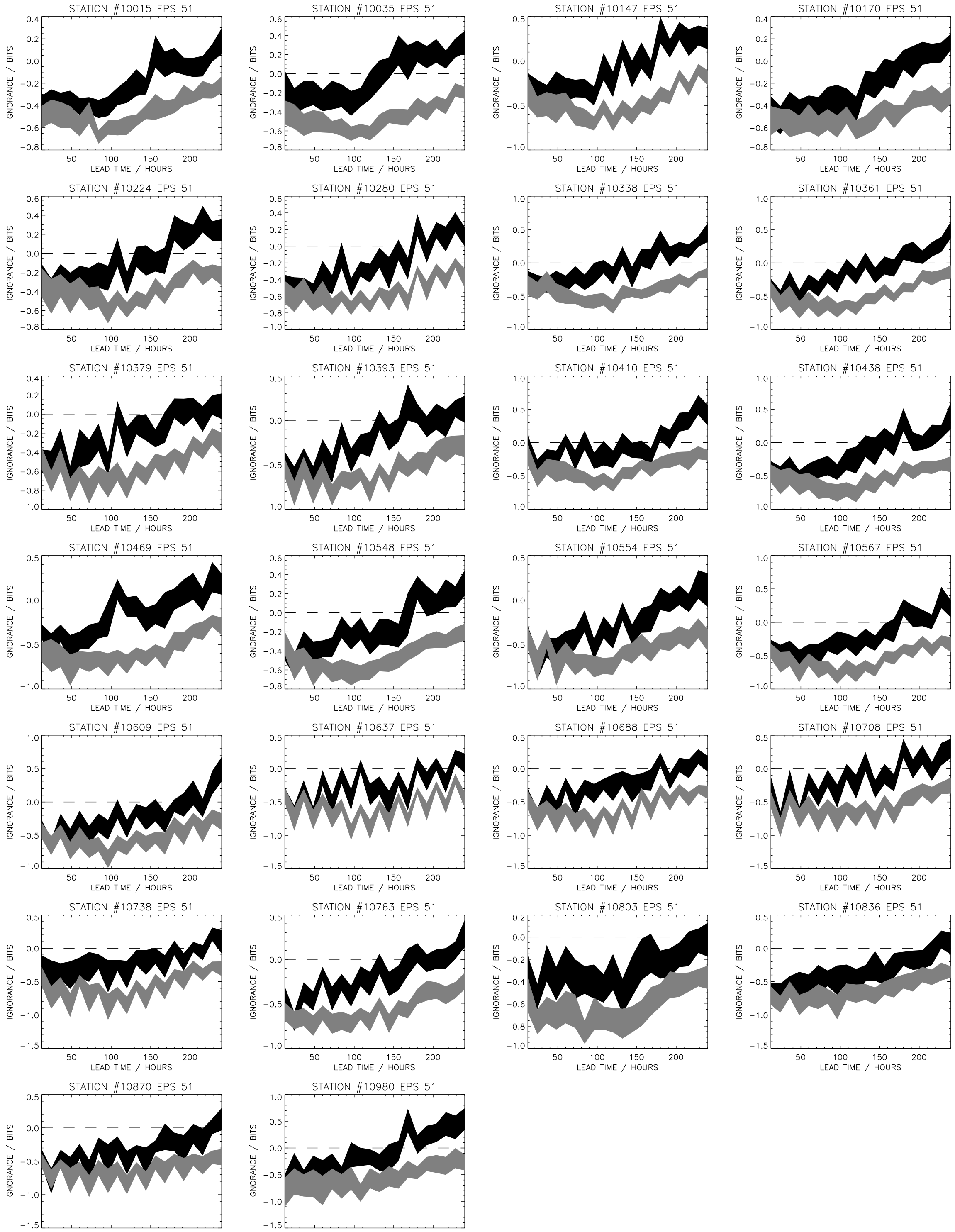


Figure 12: Average difference between logarithmic (ignorance) skill scores of forecasts and that of a climatological distribution. The best guess forecasts with a statistical ensemble are in black and the hyperensembles constructed using the full 51 member ECMWF ensemble are in grey. The dashed line represents the skill of climatology.

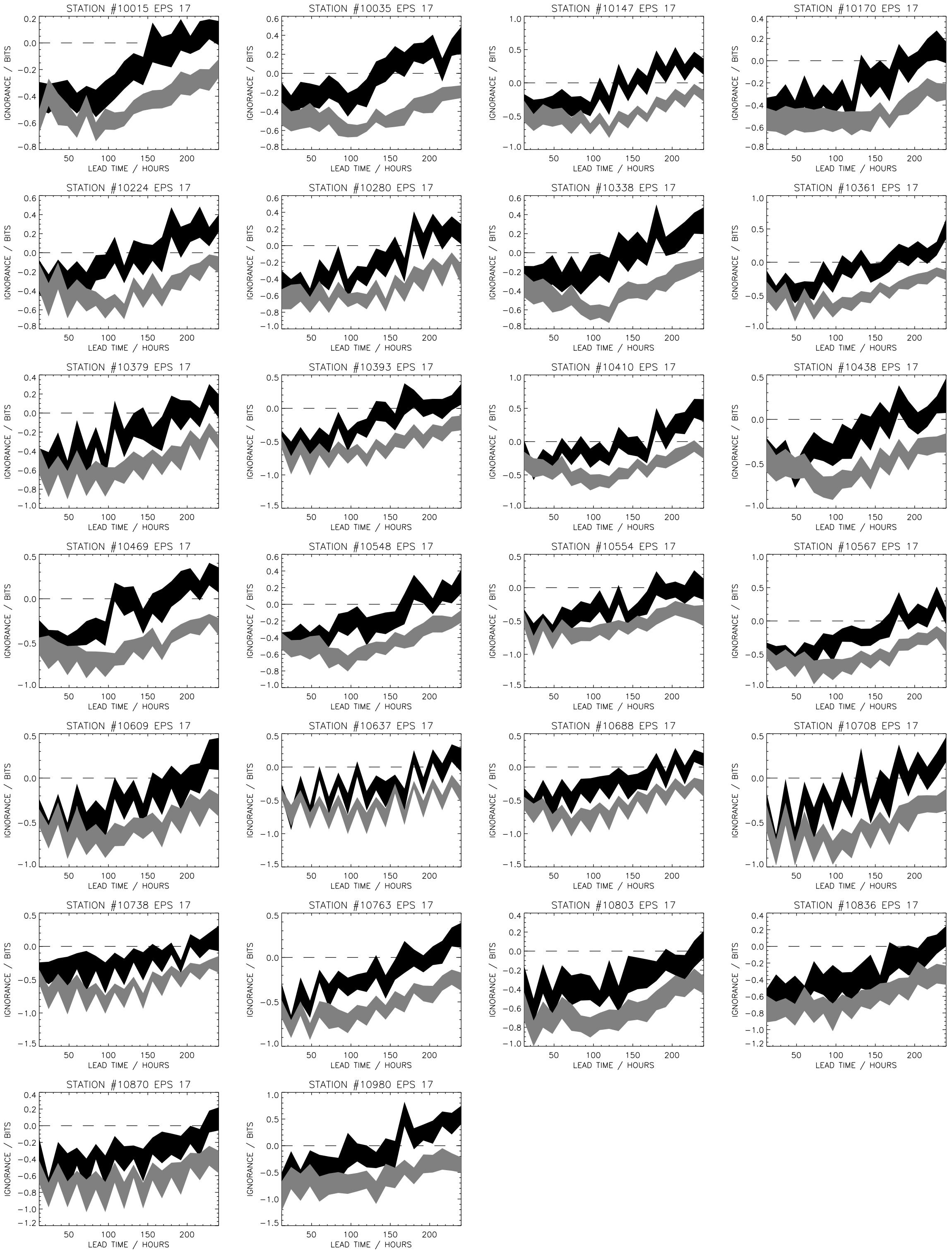
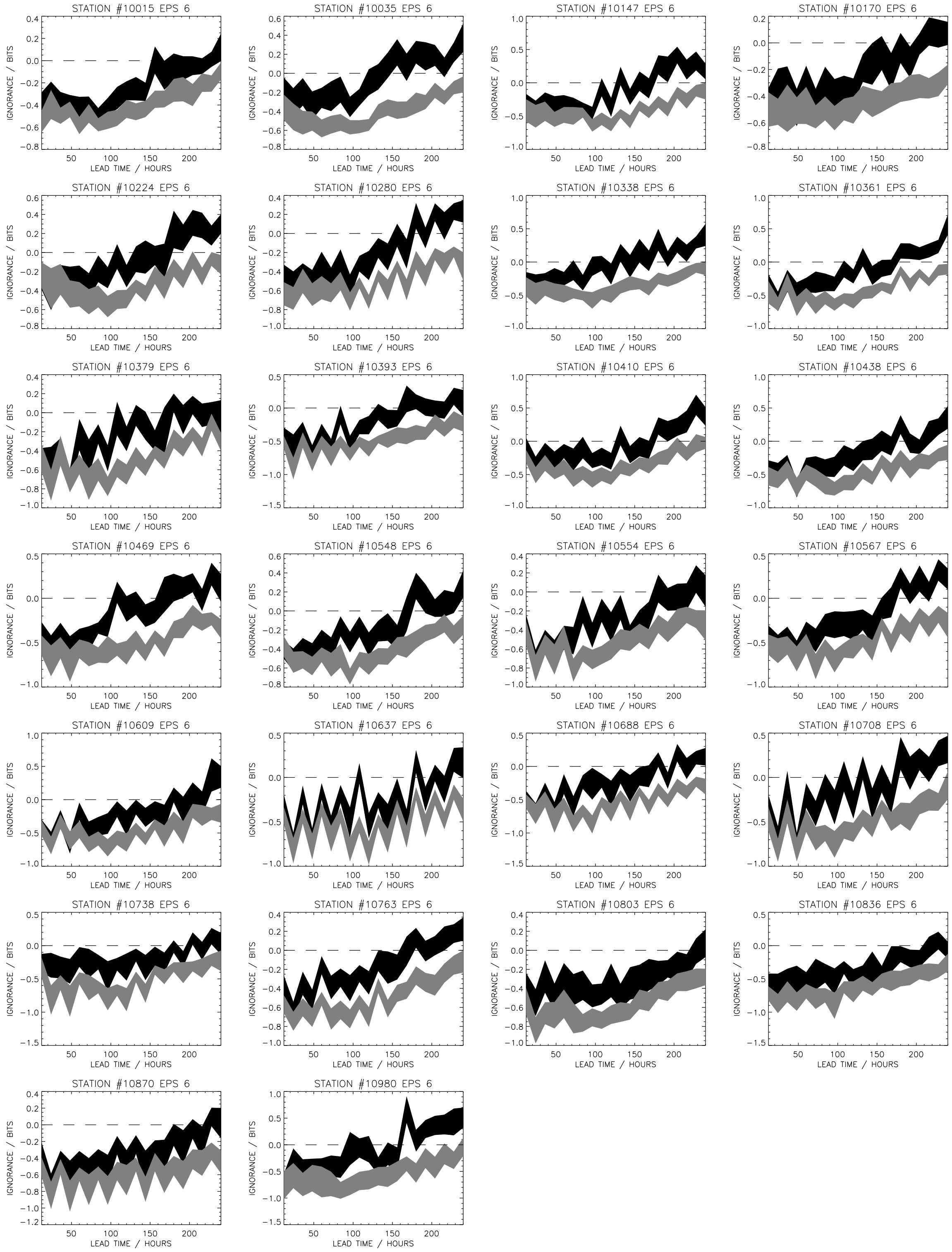


Figure 13: As Fig. 12 but the hyperensembles were constructed using only 17 members of the ECMWF dynamical ensemble.



STATION #10015 EPS 6

STATION #10035 EPS 6

STATION #10147 EPS 6

STATION #10170 EPS 6

STATION #10224 EPS 6

STATION #10280 EPS 6

STATION #10338 EPS 6

STATION #10361 EPS 6

STATION #10379 EPS 6

STATION #10393 EPS 6

STATION #10410 EPS 6

STATION #10438 EPS 6

STATION #10469 EPS 6

STATION #10548 EPS 6

STATION #10554 EPS 6

STATION #10567 EPS 6

STATION #10609 EPS 6

STATION #10637 EPS 6

STATION #10688 EPS 6

STATION #10708 EPS 6

STATION #10738 EPS 6

STATION #10763 EPS 6

STATION #10803 EPS 6

STATION #10836 EPS 6

STATION #10870 EPS 6

STATION #10980 EPS 6

Figure 14: As Fig. 12 but the hyperensembles were constructed using only 6 members of the ECMWF dynamical ensemble.

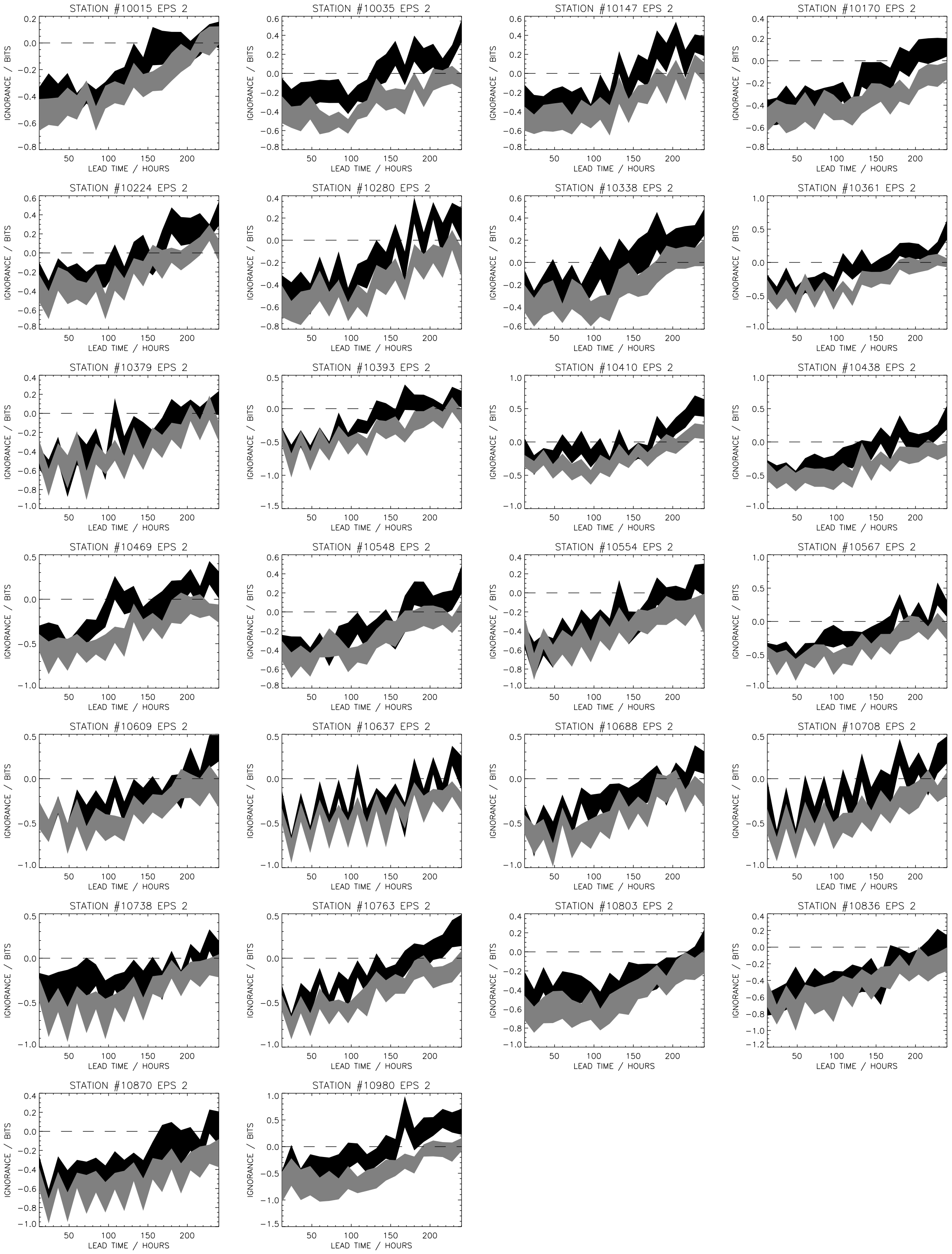


Figure 15: As Fig. 12 but the hyperensembles were constructed using only 2 members of the ECMWF dynamical ensemble.

2.4 Method of Analogues

2.4.1 Description of the Method

An alternative approach to dressing a dynamical ensemble is the "method of analogues". This method generates a climatology *conditioned on the forecast.* This method has previously been applied to ECMWF ensembles to construct probabilistic forecasts of wind energy production [Roulston et al. 2001]. The forecast is projected into a *forecast space.* Analogues are then identified by finding historical forecasts which are close to the current forecast under some metric. The observations associated with these analogues can then be treated as samples of the desired forecast probability distribution. There are many possible choices for the forecast space. The space may be defined using a deterministic forecast, an ensemble forecast or combinations of the two. In this study the method of analogues was applied to precipitation data at five German stations. The following methods for constructing probabilistic forecasts were used.

(i) **Climatology**: The precipitation for all the days in the relevant month for the reference period were used as a probability forecast.
(ii) **Raw ECMWF ensemble**: The distribution of 51 members of the ECMWF ensemble was treated as a probability forecast.
(iii) **ECMWF HIRES analogue**: The k historical forecasts for which the ECMWF high resolution precipitation forecasts were closest to the current ECMWF HIRES precipitation forecasts were chosen as analogues and used to construct a probability forecast.
(iv) **ECMWF EPS analogue**: The ECMWF ensemble forecasts were projected into a three dimensional forecast space, defined by the 10$^{\text{th}}$, 50$^{\text{th}}$ and 90$^{\text{th}}$ percentiles of the ensemble of precipitation forecasts. The k historical forecasts closest to the current EPS forecast in this space were chosen as analogues to construct a probability forecast.
(v) **DWD analogue**: As for the ECMWF HIRES analogue but the DWD medium range forecast was used instead.
(vi) **ECMWF HIRES/DWD analgogue**: As for the ECMWF EPS analogue except that the forecasts were projected into a two dimensional forecast space, defined by the ECMWF high resolution forecast and the corresponding DWD forecast.
(vii) **ECMWF HIRES-EPS/DWD analogue:** As for the ECMWF EPS analogue except that the forecasts were projected into a five dimensional forecast space, defined by the 10$^{\text{th}}$, 50$^{\text{th}}$ and 90$^{\text{th}}$ percentiles of the ECMWF ensemble forecast, the ECMWF high resolution forecast *and* the DWD medium range forecast.

In cases (iii)-(vii) then number of analogues was $k = 30$. For cases (iv), (vi) and (vii), in which the forecast space is multi-dimensional, all the coordinates defining the forecast space were normalized to have zero mean and unit variance—thus giving all coordinates equal weight in determining which historical forecasts were "close" to the current forecast.

2.4.2 Results for the Method of Analogues

To evaluate the probabilistic forecast generated using tghe methods outlined above a simple two bin quantization was used—rain, or no rain. The mean ignorance (logarithmic) skill scores for the probabilistic forecasts are shown in Figs. 16-20. The climatological forecasts have an average ignorance of about 0.97 bits. If it rained on 50% of days the climatological ignorance would be exactly 1 bit—reflecting the equiprobable outcomes of rain and no rain. In all cases the ignorance score is the difference between the ignorance score of the particular forecast and the ignorance score of the climatological forecast. A value less than zero indicates that the forecast contains more information than climatology. To obtain an estimate of the uncertainty in the estimate of the mean ignorance due to finite sampling, fifty sets of ignorance scores were generated by resampling, with replacement, from the original sample. This provided an estimate of the mean ignorance and a standard deviation in the estimate. The lines in Figs. 16-20 have thicknesses equal to two standard deviations in the estimate of the mean.

The most obvious result that can be seen from Figs. 16-20 is that using the raw ECMWF ensemble to directly construct a forecast for the probability of rain leads to *very bad* probability forecasts. In all cases, the raw ECMWF distribution is substantially worse than the climatological distribution. Using the high resolution ECMWF forecasts to find analogues provides skillful probability forecasts. Typically at a lead time of 96 hours these forecasts are between 0.2 and 0.3 bits "less ignorant" than climatology. Using the 10^{th}, 50^{th} and 90^{th} percentiles of the ECMWF ensembles to identify analogues provides slightly better probability forecasts. The difference is most noticeable at a lead time of 168 hours where the EPS analogue has an ignorance of about 0.05 bits less than the HIRES analogue. Identifying analogues based only on the DWD medium range forecast provides probability forecasts considerably more skillful than climatology, but not quite as skillful as when the ECMWF forecasts are used. Without information on changes made in the ECMWF and DWD models during the evaluation period it is not possible to say whether the difference in their performances is due to differences in the models or due to differences in the quality of the forecast error information. [3] Using a combination of ECMWF and DWD forecasts to identify analogues does not lead to a significant improvement over using only the ECMWF EPS analogues.

In general it can be said that using the raw ECMWF ensemble to construct a probability forecast leads to forecasts with less skill than using climatology, whereas, using any reasonably chosen analogues to construct probability forecasts gives a substantial improvement over climatology.

The method of analogues was also applied to windspeed data from the same five stations. The $k = 30$ closest forecasts in the space defined by the 10^{th}, 50^{th} and 90^{th} percentiles of the ECMWF ensemble were used as analogues. These analogues were then used to construct a forecast of the probability that the observed windspeed at the station would exceed the mean windspeed for that station. The results are shown in Fig. 21.

[3] All methods that use historical forecasts to estimate information about forecast errors assume that the forecast error statistics are stationary over the relevant period. The skill of forecasts generated using the method of analogues depends on the quality of error statistics, as well as on the intrinsic quality of the model. If the model is frequently changed, the nonstationarity this introduces into the historical error statistics can degrade the quality of the final forecasts—even if the model itself has been improved.

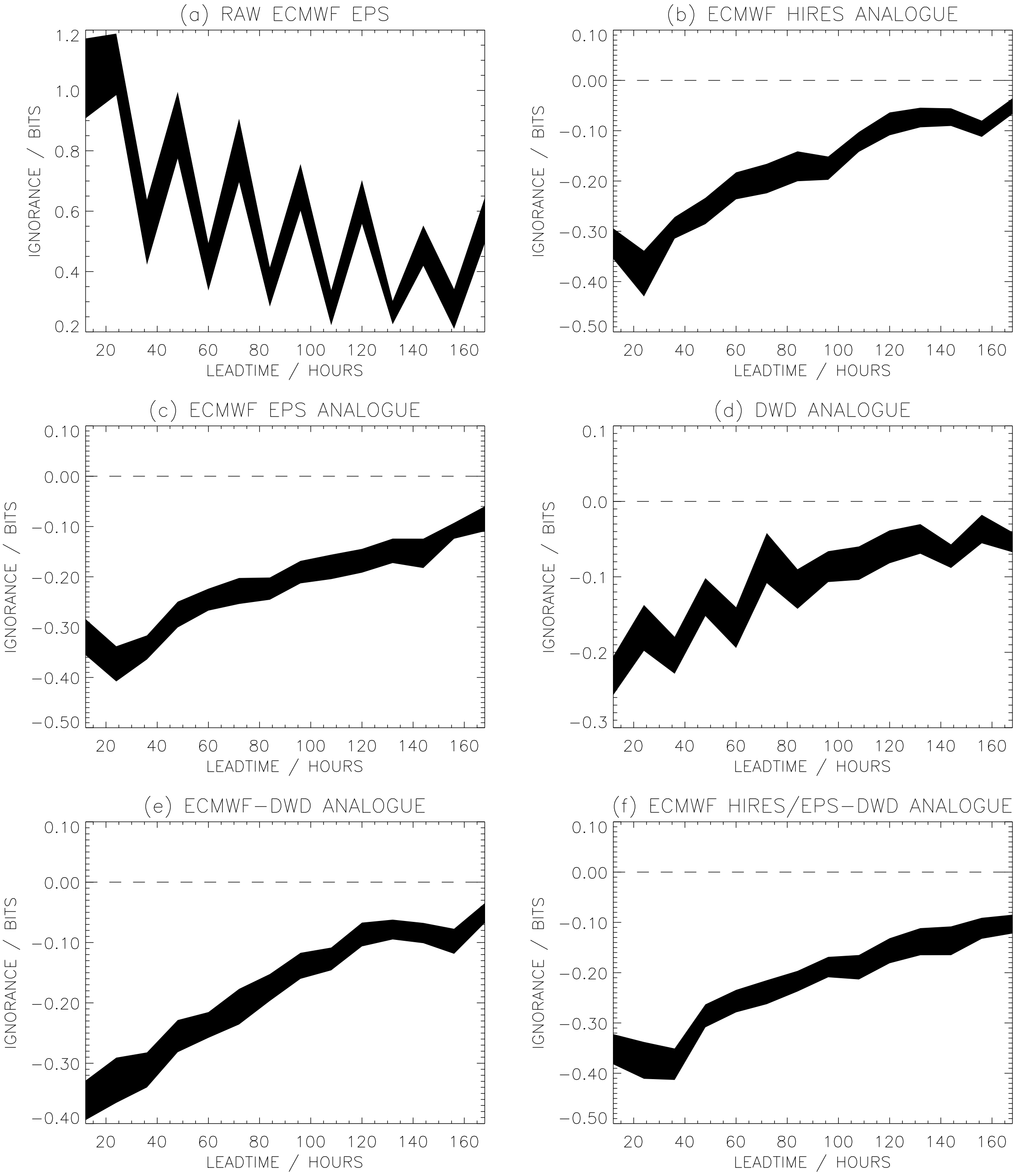


Figure 16: A comparison of ignorance scores of probabilistic precipitation forecasts constructed using different methods for station 10379. All the ignorance scores are measured relative the climatological probabilistic forecast. A score less than zero indicates greater skill than climatology. The thickness of the lines are two standard deviations in the estimate of the mean obtained using a bootstrap resampling method.

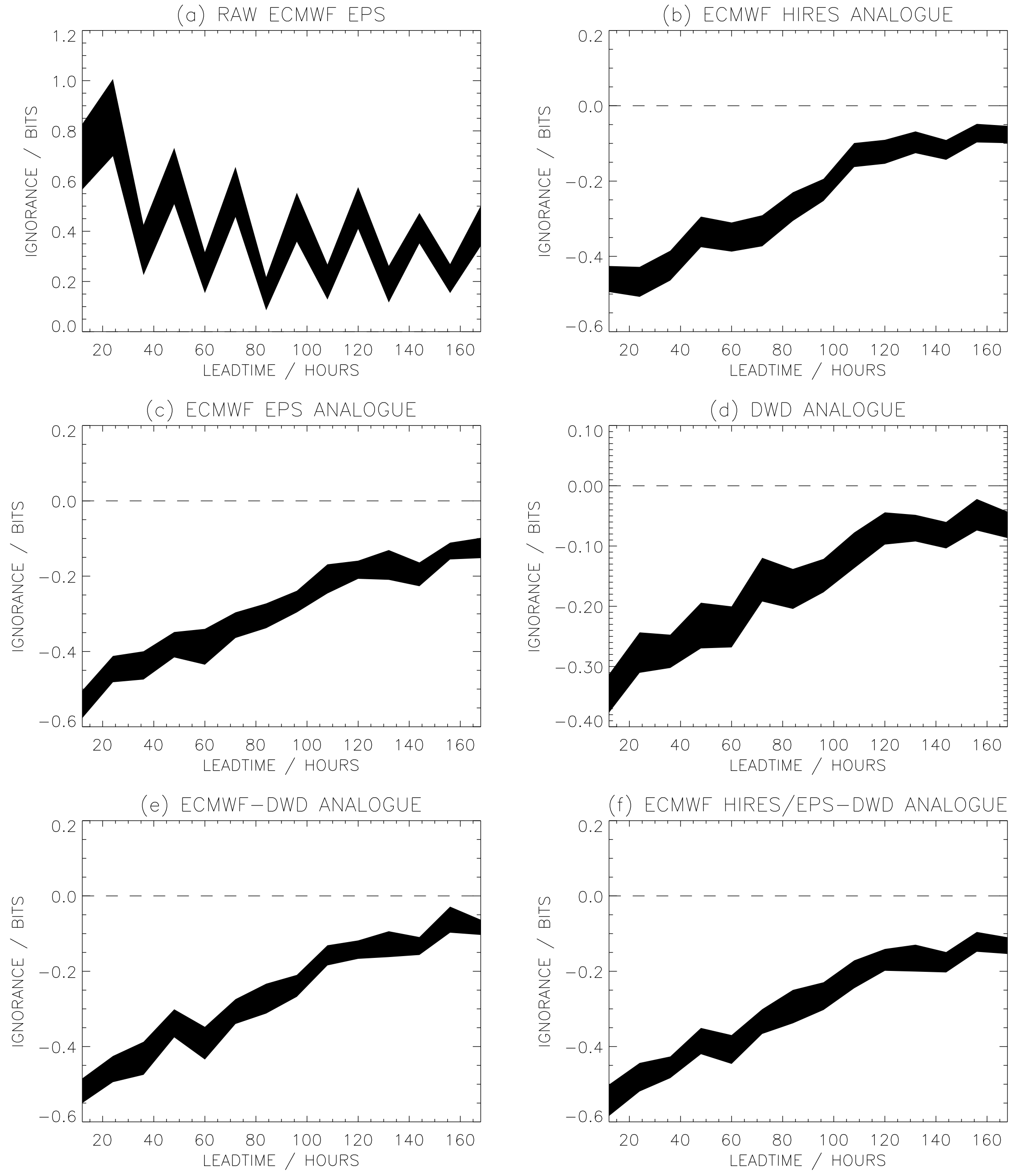


Figure 17: As Fig. 16 but for station 10410.

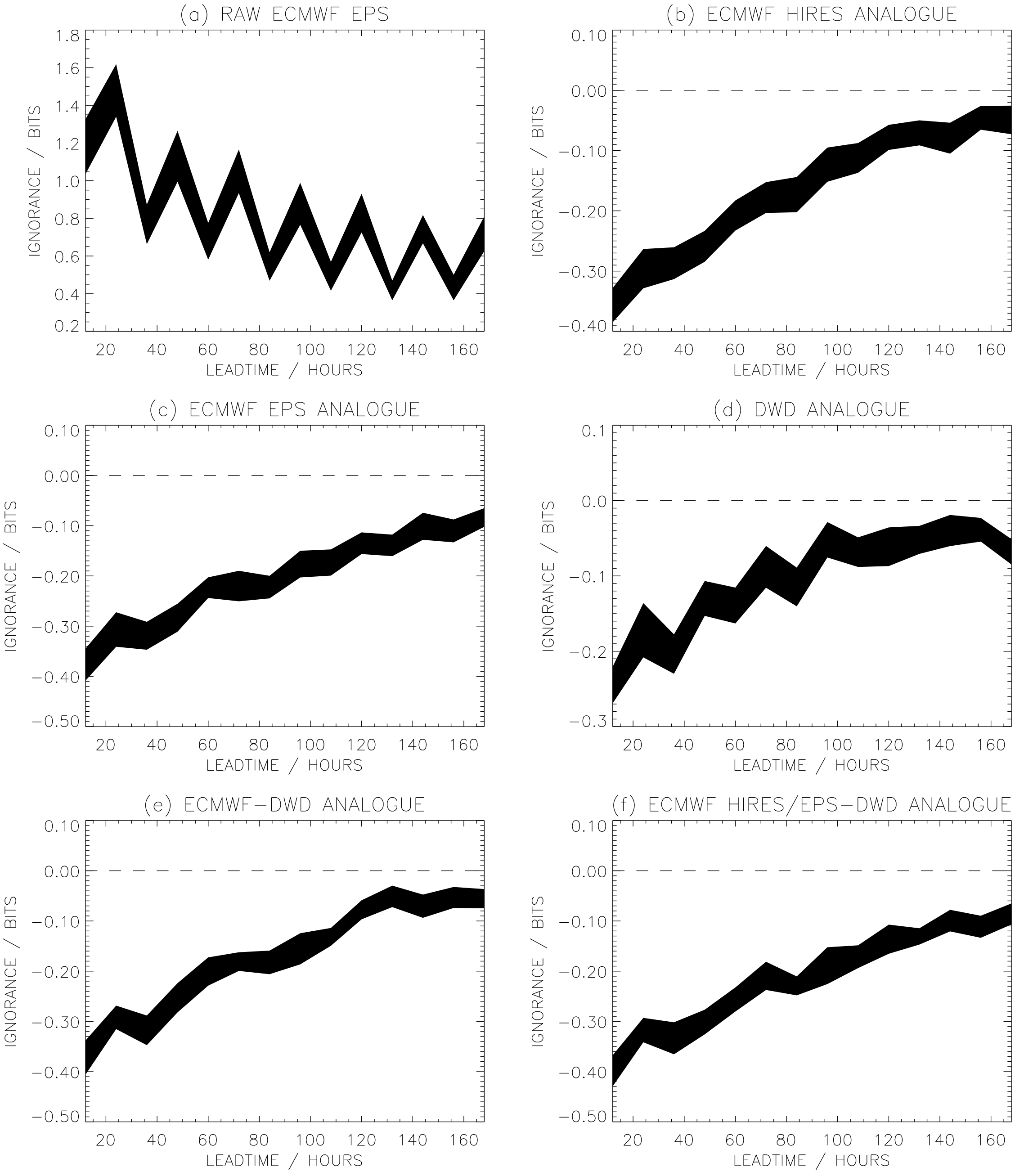


Figure 18: As Fig. 16 but for station 10469.

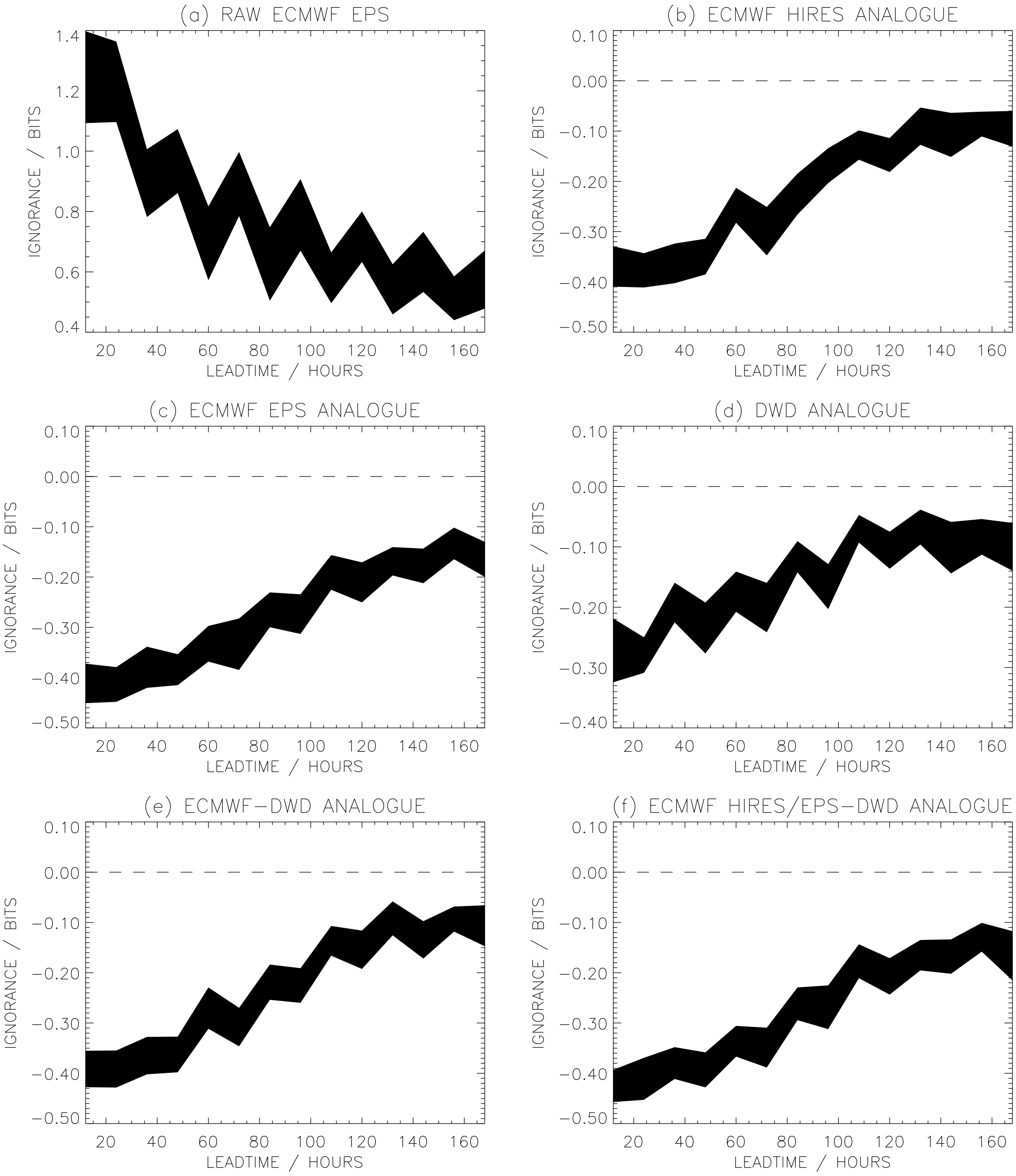


Figure 19: As Fig. 16 but for station 10637.

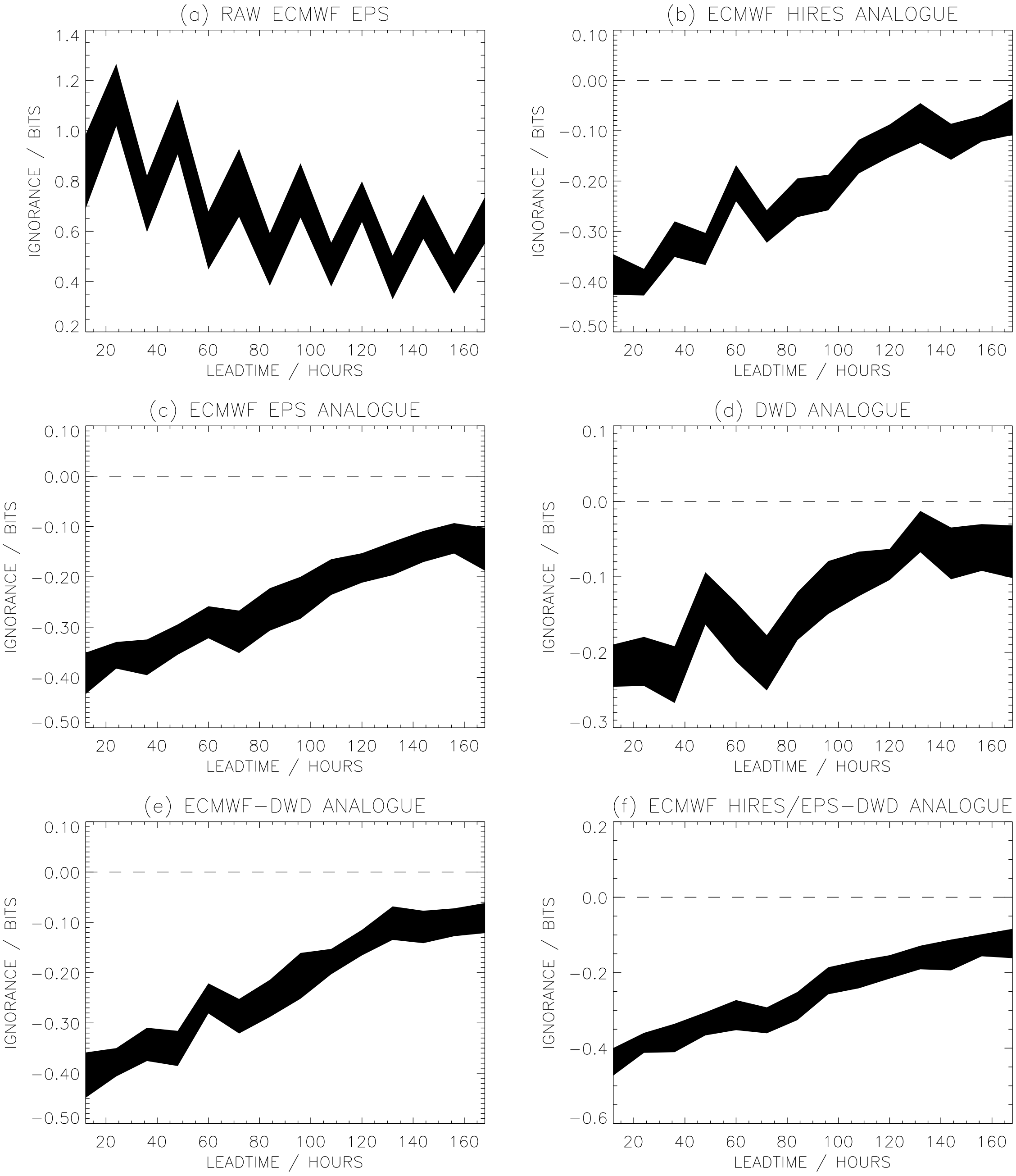


Figure 20: As Fig. 16 but for station 10738.

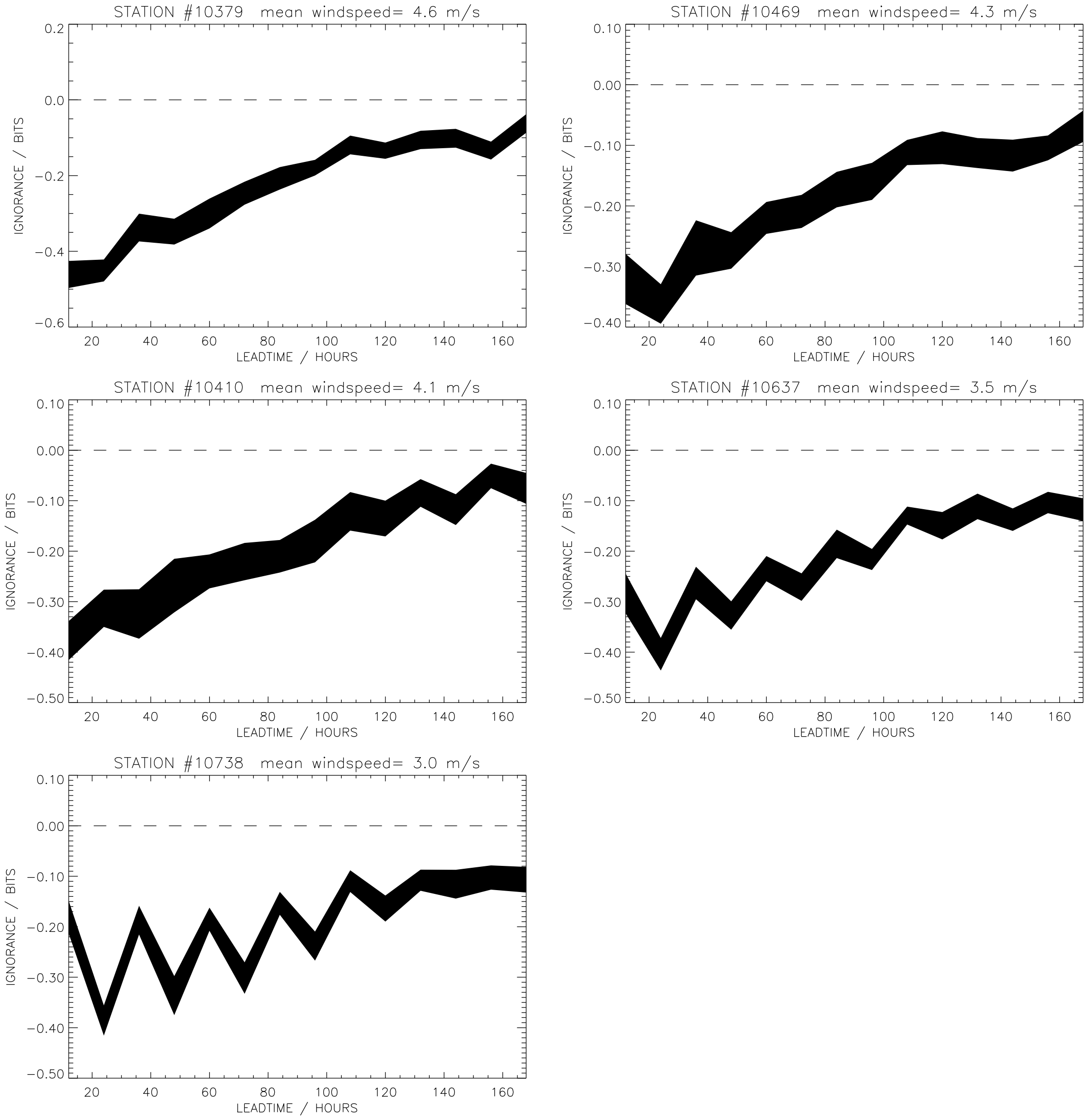


Figure 21: The average ignorance (with respect to climatology) of probabilistic windspeed forecasts constructed using the method of analogues, where analogues were found in the space of the 10^{th}, 50^{th} and 90^{th} percentiles of the ECMWF EPS forecast. The probabilistic forecast was the probability of the windspeed exceeding the mean windspeed at that station.

FORECAST QUALITY	FRACTION OF HYPERENSEMBLE MEMBERS WITHIN 3°C OF MEAN
GOOD	greater than 80%
MEDIUM	between 50% and 80%
BAD	less than 50%

Table 1: Temperature forecast quality classes.

2.5 Forecast Classes from Probabilistic Forecasts

2.5.1 Forecast Quality Classification

The ideal forecast is an estimate of the probability distribution of future outcomes. Sophisticated users, who can quantify the weather dependence of their utility to a reasonable degree, can use fully probabilistic forecasts to make choices that maximize their expected utility. Less sophisticated users may still obtain value from a crude estimate of the quality of a traditional, deterministic forecasts. Estimates of forecast quality can be made using the hyperensembles generated using the method of best members. To demonstrate how this can be done the mean of a hyperensemble generated around the ECMWF EPS dynamical ensemble was treated as a deterministic forecast. The quality of the forecast was then estimated *a priori* by counting the fraction of hyperensemble members that came within 3°C of the hyperensemble mean. The forecast quality classes were defined according to Table 1.

Figure 22 shows the success rate for forecasts according to their forecast quality class. It can be seen that, while the forecast quality class does provide an indicator of the success rate, the system is not strictly *reliable*. If the classification were reliable all the high skill forecasts would lie at, or above, 80% and all the low skill forecasts would lie at, or below, 50%, with the medium skill forecasts lying between 50% and 80%. This is approximately true for many stations but seriously violated for for station #10015. Note also that the quality of a forecast with a particular skill class *is not dependent on forecast lead time.* The reduction in predictability at longer lead times is reflected by the smaller number of good forecasts, and the greater number of bad forecasts. The quality of good and bad forecasts remains approximately constant at all lead times. The result in Fig. 22 can be contrasted with the results of temperature forecast skill prediction obtained using ensemble mode population [Ziehmann 2000; Ziehmann 2001]. Using that method even forecasts classed as having high predictability only have a success rate of 70%. [4]

[4] This result used 10 climatologically equiprobable bins and success was defined when the observation fell into the same bin as the forecast. The typical bin size was generally less than the 3°C tolerance used in the present work. This difference could account for the less confident forecasts obtained using ensemble mode population as a predictor of forecast skill.

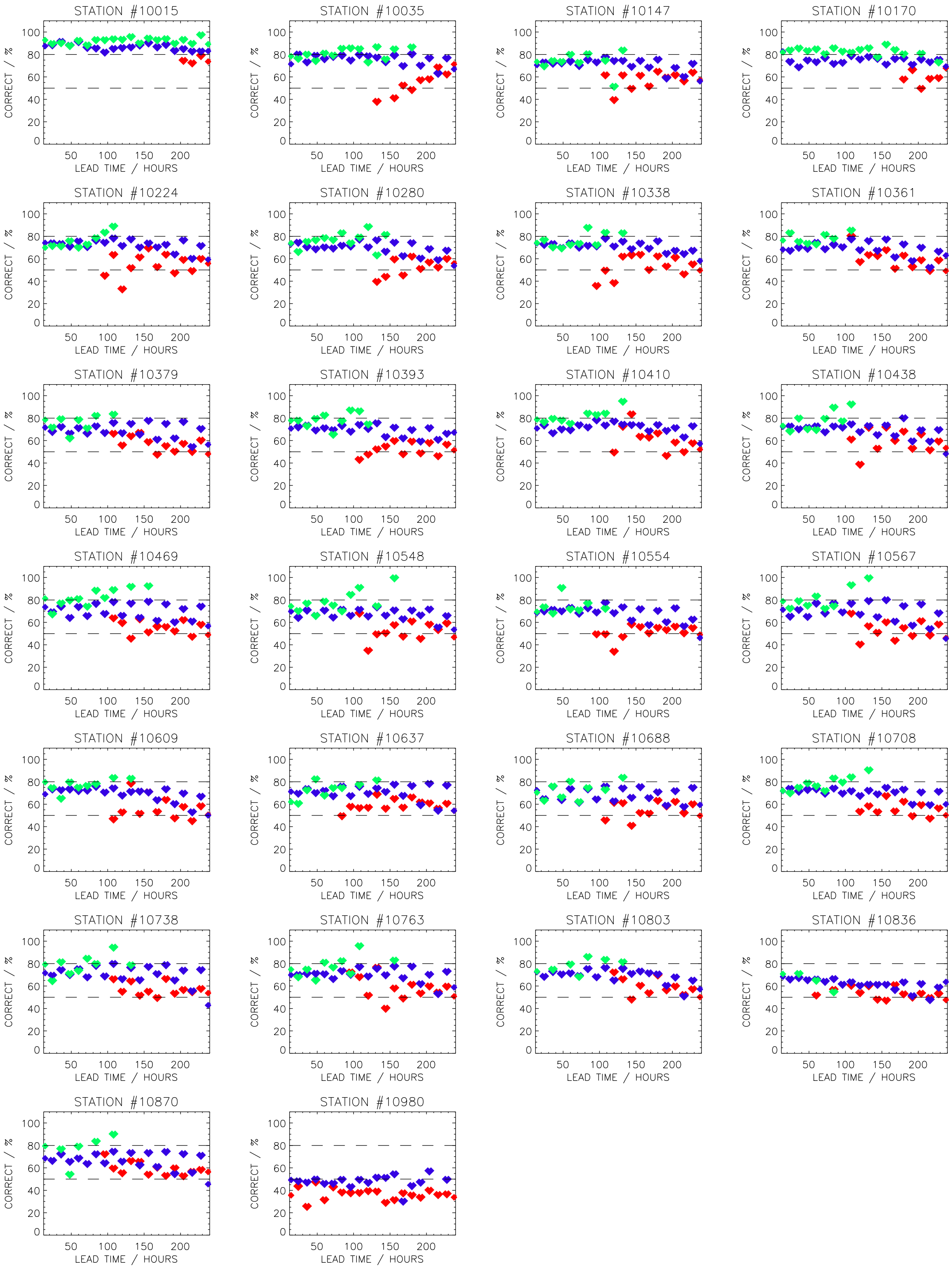


Figure 22: The percentage of correct temperature forecasts (within 3°C of the forecast value) for different lead times and different a priori skill classes (green=good, blue=medium, red=bad). Only forecast classes for which there were 10 or more cases are shown.

FORECAST CLASS	FRACTION OF ANALOGUES WITH PRECIPITATION > 0
HIGH PROBABILITY	greater than 80%
MEDIUM PROBABILITY	between 30% and 80%
LOW PROBABILITY	less than 30%

Table 2: Precipitation forecast probability classes.

2.5.2 Forecast Probability Classification

Temperature is a continuous quantity, so a value of the expected temperature and an estimate of the quality of this forecast is required. For a simpler forecast question, such as “will it rain?”, a classification system can be designed which is based directly on the probability of the event occuring (rather than on the probability of a forecast being right). To illustrate this, the probability forecasts of precipitation, constructed by finding analogues in the forecast space of the 10^{th}, 50^{th} and 90^{th} percentiles of the ECMWF ensemble, were converted into forecast classes according to Table 2.

Figure 23 shows the results of the precipitation forecast classification for five stations. It can be seen that the classification system is reasonably reliable. The high probability forecasts always have a precipitation frequency greater than 80%. The low probability forecasts typically have precipitation frequencies less than 30% although occasionally they reach 40%. The medium probability forecasts have frequencies which, except in a few cases, lie between 30% and 80%. As with the temperature forecasts, the decrease in predictability at longer lead times is reflected by a change in the relative numbers of different forecast classes. Few high probability forecasts are issued at lead times exceeding 80 hours. However, when such forecasts are issued at longer lead times they are almost as dependable as the ones issued at very short lead times.

The results in Fig. 23 should be contrasted with those in Fig. 5. Using the method of analogues it is possible to issue probability forecasts of rain exceeding 80% which are reliable, whereas, using ensemble mode population as a predictor of forecast skill it is not really possible to issue a forecast of rain with more than about 70% confidence. [5]

[5] Since the same binary rain/no rain binning was used to define success in Figs. 23 and 23, the qualification concerning the definition of success that applies to the temperature result does not apply to the precipitation result.

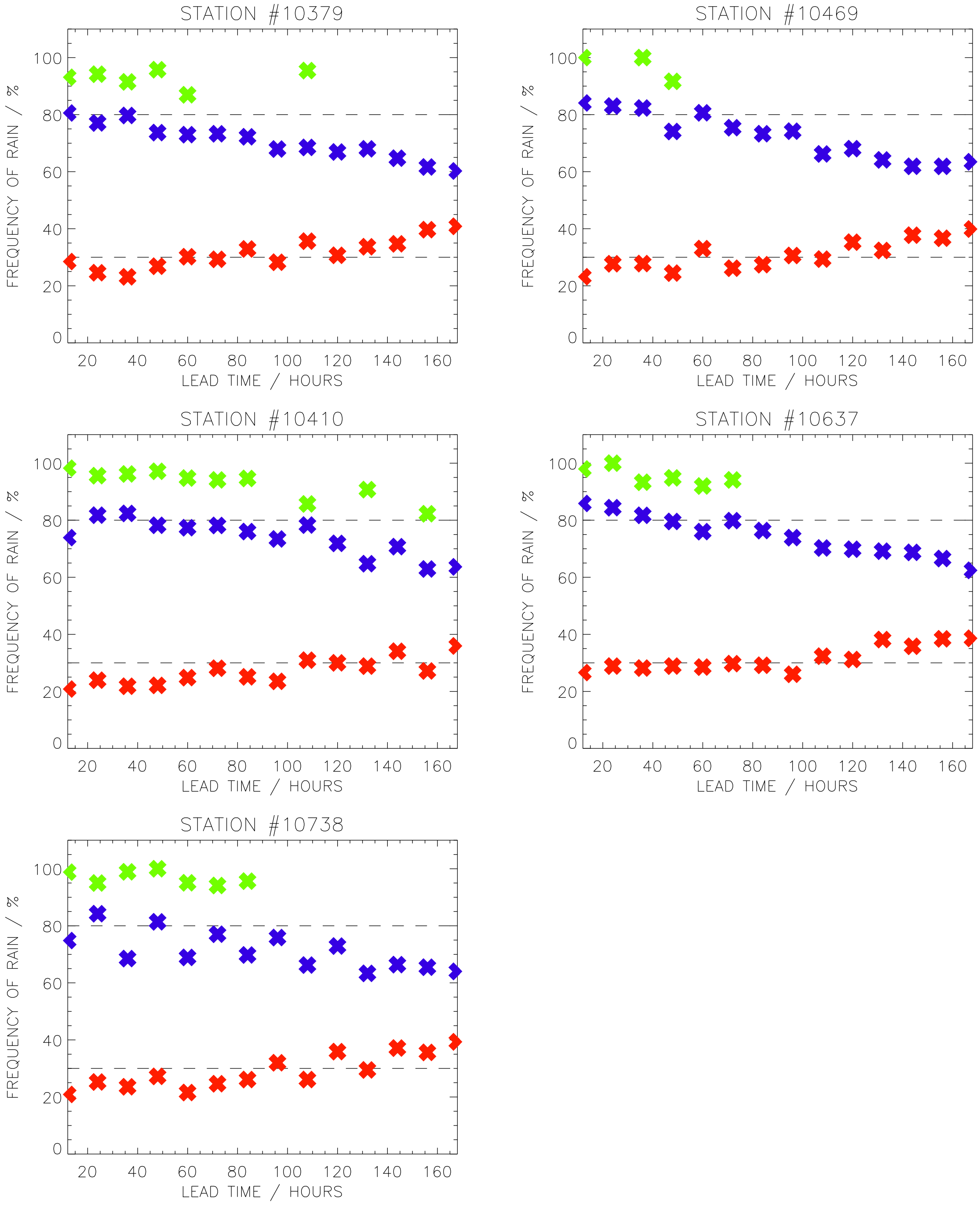


Figure 23: The percentage of times with nonzero precipitation for different precipitation forecast probability classes (green=high, blue=medium, red=low). Only forecast classes for which there were 10 or more cases are shown.

3 Conclusions

In this study two general approaches for constructing predictors of forecast skill have been presented. The first approach was to use a single statistic of the ensemble distribution (mode population or entropy) as a predictor of the forecast skill. The second approach was to attempt to generate a probabilistic forecast using information contained in the ensemble forecast and/or a deterministic forecast. A predictor of forecast skill could then be defined based on the probabilistic forecast. For the locations in Germany that were studied, both approaches were found to produce useful predictors of forecast skill. Constructing skill predictors from probabilistic forecasts, however, allowed more confident forecasts to made. The probabilistic forecasts also offer more flexibility in the design of forecast skill predictors.

The probabilistic forecasts for temperature, constructed using the ECMWF ensemble prediction system, and the method of "best member dressing", were found to have skill out to a lead time of 10 days. "Skill", in this sense, meaning that, on average, they contained more information than climatological probabilities, and outperformed climatological probabilities when either a quadratic (Brier) or logarithmic (ignorance) scoring rule was used to evaluate them. It was also found that using as few as 6 EPS ensemble members did not lead to a major degradation in the information content of the probabilistic forecast when temperature at a single location was the target forecast variable. This may not be the case when a multivariate forecast is required. Probabilistic precipitation forecasts, constructed from the ECMWF ensemble prediction system, using the "method of analogues", were found to have skill out to a lead time of 7 days.

Both the probabilistic temperature and precipitation forecasts were used to construct predictors of forecast skill. In both case the probabilistic forecasts enabled forecasts to be classed according to confidence. The skill predictors for the precipitation forecasts were found to be the most reliable—this may be due to the superiority of the "method of analogues" rather than due to intrinsic differences between forecasting temperature and precipitation.

Appendix: Skill scores and cost-loss

The cost-loss score is the realized utility of a user attempting to maximize their expected utility. If the user's cost-loss matrix for a binary event is:-

	EVENT HAPPENS	EVENT DOESN'T HAPPEN
USER ACTS	C	C
USER DOESN'T ACT	L	0

where C is the cost of acting, and L is the loss incurred if action is not taken and the event occurs. If the forecast probability of the event is f, then to minimize their expected loss the user should act if $p \geq C/L$. If the actual probability of the event is p, then the expected loss of the user, U, will be

$$U = \begin{cases} C & \text{when} \quad f \geq C/L \\ pL & \text{when} \quad f < C/L \end{cases} \tag{11}$$

If $u(\alpha)d\alpha$ is the number of users with a cost-loss ratio of $\alpha = C/L$, then the expected loss, averaged over users and normalized in units of L, is

$$\langle U \rangle = \int_0^f \alpha u(\alpha) d\alpha + p \int_f^1 u(\alpha) d\alpha \tag{12}$$

Differentiation w.r.t. f gives

$$\frac{d\langle U \rangle}{df} = u(f)(f - p) \tag{13}$$

The expected quadratic (Brier) score is given by

$$\begin{aligned} \langle BS \rangle &= p[(1-f)^2 + (1-f)^2] + (1-p)[f^2 + (1-(1-f))^2] \\ &= 2p - 4pf + 2f^2 \end{aligned} \tag{14}$$

Differentiation of Eq. 14 w.r.t. f gives

$$\frac{d\langle BS \rangle}{df} = 4(f - p) \tag{15}$$

A comparison of Eqs. 13 and 15 shows that the expected Brier score is linear with the average expected utility if $u(\alpha)$ is uniform. The expected logarithmic (ignorance) score is given by

$$\langle IGN \rangle = -p \log f - (1-p) \log(1-f) \tag{16}$$

Differentiation gives

$$\frac{d\langle IGN \rangle}{df} = \frac{f - p}{f(1-f)} \tag{17}$$

Comparing Eqs. 13 and 17 indicates that the expected ignorance is linear with the average expected utility if $u(\alpha) \propto [\alpha(1-\alpha)]^{-1}$. The singularities in the distribution at $\alpha = 0$ and $\alpha = 1$ arise because a user with $\alpha = 0$ would always act, and a user with $\alpha = 1$ would never act, thus there is no decision making scenario for these values.

References

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

Buizza, R., 1997: Potential forecast skill of ensemble prediction and spread skill distribution of the ECMWF ensemble prediction system, *Mon. Wea. Rev., 125,* 99-119, 1997.

Buizza, R., Miller, M. and Palmer, T.N., 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system, *Quart. J. Roy. Met. Soc., 125,* 2887-2908.

Efron, B. and Tibshirani, R., 1986: Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, *Statistical Science, 1,* 54-77.

Katz, R.W. and Murphy, A.H., 1987: Quality/value relationships for imperfect information in the umbrella problem, *The American Statistician, 41,* 187-189.

Katz, R.W. and Murphy, A.H., 1997: Forecast value: prototype decision-making models, in *Economic Value of Weather and Climate Forecasts* (eds. Katz and Murphy), 183-217, Cambridge Univ. Press., Cambridge.

Lindley, D.V., 1985: *Making Decisions,* John Wiley and Sons, London.

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

Murphy, A.H., 1966: A note on the utility of probabilistic predictions and the probability score in the cost-loss ratio decision situation. *J. App. Meteor., 5,* 534-537.

Palmer, T.N., 2000: Predicting uncertainty in forecasts of weather and climate, *Rep. Prog. in Phys., 63,* 71-116.

Palmer, T.N., 2001: A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parameterization in weather and climate prediction models, *Quart. J. Roy. Met. Soc., 127,* 279-304.

Richardson, D.S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system, *Quart. J. Royal Met. Soc., 126,* 649-667.

Richardson, D.S., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size, *Quart. J. Royal Met. Soc. (in press)*

Roulston, M.S. and Smith, L.A., 2001: Evaluating probabilistic forecasts using information theory, *Mon. Wea. Rev.* (to appear)

Roulston, M.S. and Smith, L.A., 2001: Combining dynamical and statistical ensembles, submitted to *Tellus A*

Roulston, M.S. and Smith, L.A., 2001: Statistical ensembles, dynamical ensembles and hybrid ensembles, *American Geophysical Union Fall Meeting 2001 (abstract NG42A-0408)*

Roulston, M.S., Kaplan, D.T., Hardenberg, J. and Smith, L.A., 2001: Value of the ECMWF ensemble prediction system for forecasting wind energy production, *European Wind Energy Conference 2001 (abstract OD3.5)*

Smith, L.A., Roulston, M.S. and Hardenberg, J., 2001: End to End Ensemble Forecasting: Towards evaluating the economic value of the ensemble prediction system, *ECMWF Technical Memorandum No. 336 2001.*

Toth, Z., Zhu Y. and Marchok, T., 2001: On the ability of ensembles to distinguish between forecasts with small and large uncertainty, *Weather Forecasting*

Ziehmann, C., 2000: Skill prediction of local weather forecasts based on the ECMWF ensemble, *DWD technical report*

Ziehmann, C., 2001: Skill prediction of local weather forecasts based on the ECMWF ensemble, *Nonlinear Processes in Geophysics* (in press)