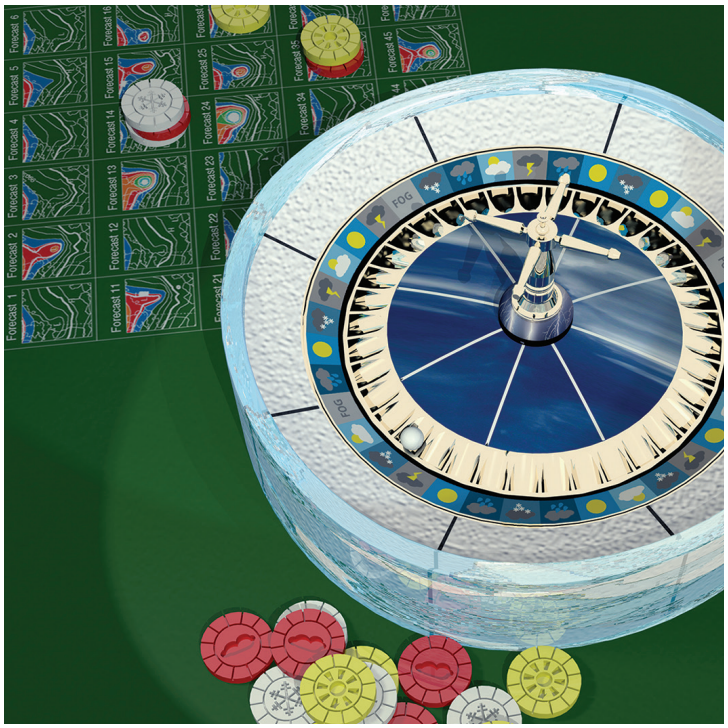
ECMWF Feature article

from Newsletter Number 106 – Winter 2005/06

METEOROLOGY

Comparing and combining deterministic and ensemble forecasts: How to predict rainfall occurrence better



www.ecmwf.int/en/about/news-centre/media-resources

doi:10.21957/cd347812th

This article appeared in the Meteorology section of ECMWF Newsletter No. 106 – Winter 2005/06, pp. 17–23.

Comparing and combining deterministic and ensemble forecasts: How to predict rainfall occurrence better

Mark J. Rodwell

Present ECMWF medium-range forecast products include a high-resolution (T511) "deterministic" forecast (termed here "DET"), a lower-resolution (T255) "control" forecast (termed here "CNT") and a 50-member T255 "ensemble prediction system" ("EPS") of forecasts initiated from perturbed initial conditions. Naturally, there is considerable debate about which forecast system, DET or EPS, is better, or at least about how ECMWF should divide its resources between deterministic and probabilistic forecasts. Here these two systems are compared using deterministic and probabilistic skill scores but the main conclusion is that direct comparison is not straightforward and is actually not very useful. This led to a different approach being taken which attempts to harness the best elements of DET and EPS to produce a combined prediction system ("CPS") of European station-location precipitation.

Rainfall has fine spatial scale structure, low predictability and is important to forecast and thus offers perhaps the best chance of finding benefit in both forecast systems which is relevant to the users. It is found that the CPS is significantly better than EPS at forecasting the probability of occurrence of European rainfall at all lead-times to D+10. The optimal weight applied to the DET forecast within the CPS is found to be equivalent to 17 EPS members at day+1 (D+1), dropping to 2.5 EPS members by D+10. It is found that the aspect of DET that leads to the increased skill is its higher resolution, and not the fact that it is initiated from a slightly better (unperturbed) estimate of the true state of the atmosphere. Results point to the variable resolution EPS (VAREPS) as the optimal framework for precipitation forecasting in that the benefits of short lead-time resolution are combined with longer lead-time probabilities.

Definition of parameters and recent improvements in deterministic forecast skill

Clearly it is important to know how good our forecasts are and to document our progress over the years in improving these forecasts. The parameter most commonly used to score weather forecasts is geopotential height at 500 hPa (Z_{500}). Spatial anomaly correlations between the forecast and observed (i.e. analysed) Z_{500} have improved over the years so that, for the Northern Hemisphere as a whole, a D+7 forecast made in 2001 was on average as good as a D+5 forecast made in 1980 (*Simmons & Hollingsworth,* 2002). Here we consider how the skill of Z_{500} and two other parameters, including precipitation, have changed since 2001.

The top set of curves in Figure 1 show annual means of the spatial Anomaly Correlation Coefficients (ACCs) of the DET forecasts for European Z_{500} as a function of forecast lead-time. Different colours represent different years. It can be seen that the trend to improved forecast skill (up to D+6) has continued year-on-year. Beyond D+6 the trend is less clear but this may simply reflect increased uncertainty in the ACC estimate at longer lead-times, as indicated by the 95% confidence intervals in the plot.

Although there has been a continual improvement in the prediction of Z_{500} over the years it is clear that, for short forecast lead-times, there is little extra skill in terms of the anomaly correlation to be gained in the future. However, this does NOT mean that our job is complete as far as extratropical medium-range forecasting is concerned. The Z_{500} field emphasises the atmospheric circulation on very large spatial scales (perhaps 1000 km). It is clearly worth investigating the forecast skill on shorter spatial scales and using variables that are of particular interest to users of forecasts. Here, we concentrate on the skill of the DET forecast to predict two additional variables: potential temperature on the PV = 2 surface (see below) and total precipitation (convective plus stratiform).

Potential temperature on the PV = 2 surface (θ_2) is chosen because it is related to Potential Vorticity (PV) which varies on smaller spatial scales. The conservation properties of PV may allow us, in the future, to investigate the causes of forecast error. The PV = 2 surface approximates the tropopause in the extratropics. An additional advantage of using θ_2 is that it is archived from both the EPS and DET forecasts.

The middle set of curves in Figure 1 show the European ACC scores for θ_2 from the DET forecast. These scores are lower than those for Z_{500}. The trend over the years is still upwards but it can now be seen that we still have plenty of scope for improving the ACC skill of forecasts of synoptic and smaller-scale features such as tropopause folds and intense cyclones.

doi:10.21957/cd347812th

Precipitation is one of the hardest quantities to predict yet it is something that is of particular interest to users. Here we bi-linearly interpolate forecast precipitation from the model grid to European SYNOP station locations so that we are scoring our ability to predict rainfall at a point (literally raindrops falling on your head!). The station-location point precipitation is referred to here as P_p.

Annual-mean rainfall varies strongly from place to place. For example it rains much more in mountainous regions than on the plains. To ensure that our anomaly correlation score reflects our ability to predict precipitation everywhere and not just in mountainous regions, observed and predicted precipitation is divided by the monthly-mean climatological value for each station location and referred to here as $P_{p/c}$. The ACC is based on $P_{p/c}$. For numerical reasons, only stations for which the monthly-mean climatological precipitation exceeds 4 mm are considered. Between 300 and 400 stations are used on any given day.

Assessing forecast skill averaged over SYNOP stations will give more prominence to regions where the station network density is largest. There exists at ECMWF a gridded precipitation analysis based on a high spatial resolution rain gauge network. This precipitation analysis is used in verification and routine diagnostics. Using it here would avoid the network inhomogeneity issue but this would be at the expense of the skill-at-a-point feature that users are clearly interested in. For the future development of this investigation, it is intended to divide each SYNOP station's contribution to a skill score by the station network density in its vicinity. In practice, this may not lead to very different results to those quoted here but it would take account of network inhomogeneity and yet retain the desired skill-at-a-point feature.

The bottom set of curves in Figure 1 show the European ACC scores for $P_{p/c}$. Scores at D+*n* refer to the forecast of precipitation accumulated over the preceding 24 hours: D+(*n*–1) to D+*n*. As anticipated, the scores are even lower but again the trend over the years is for improvements in our prediction of rainfall. Notice that even at D+1 for the most recent year, the ACC for $P_{p/c}$ barely reaches 0.6. The value of 0.6 is often taken as the threshold for usefulness for large-scale flows although it is unclear whether the same threshold applies for precipitation. (Usefulness is discussed later in terms of probabilistic scores.) This deterministic score is complimentary to the "True Skill Score" (TSS) because the TSS involves the definition of a rainfall threshold. One of the desirable properties of such a deterministic score is that it is not complicated by changes in the tuning of ensemble spread and thus it provides a more direct method of assessing the impact on precipitation of changes to the model physics or resolution.

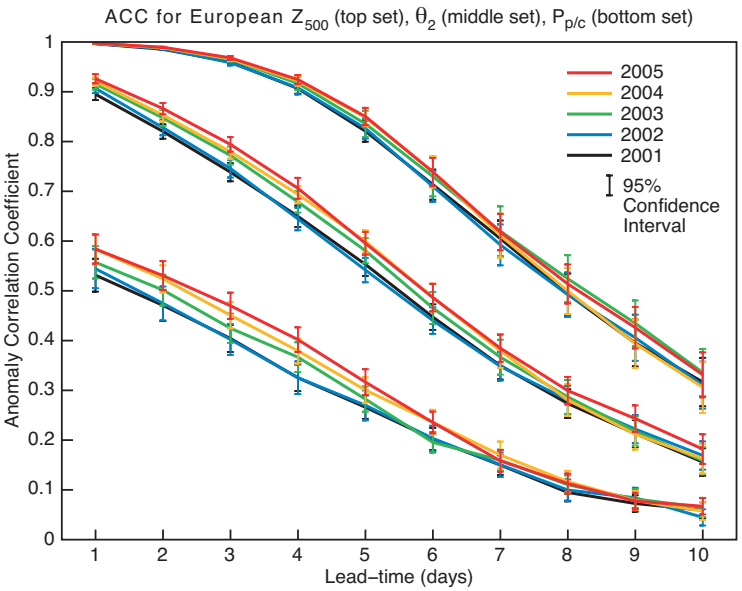


Figure 1 Spatial Anomaly Correlation Coefficients (ACCs) of European 500 hPa geopotential heights (Z_{500}, top set of curves), Potential Temperature on the Potential Vorticity (PV) surface of 2 PV units (θ_2, middle set of curves) and station-location point precipitation divided by climatology ($P_{p/c}$, bottom set of curves). Each curve is the average ACC for an entire year of 12 UTC high-resolution (T511) deterministic forecasts (termed "DET" forecasts). Different colours refer to different years. 95% two-sided confidence intervals, which take autocorrelation into account, are displayed. Z_{500} and θ_2 data are interpolated to a 2.5° regular grid before the ACCs are calculated. The climatologies used to calculate the ACCs come from the entire ECMWF 40-year re-analysis (ERA-40) for Z_{500} and θ_2 and from a 1971–2000 observational database for $P_{p/c}$. Europe is defined in this article as the region 12.5°W–42.5°E, 35.0°N–75.0°N.

doi:10.21957/cd347812th

Deterministic comparison of deterministic and ensemble prediction systems

A deterministic score requires as input a single prediction, not a probabilistic prediction. To get a single prediction from the EPS, the natural thing to do is, perhaps, to take the mean of all 50 ensemble members (referred to here as "EMN"). Figure 2 compares ACCs from the DET forecast (red) with those of the EMN (black).

- For Z_{500} (top two curves) the high-resolution forecast is superior to the ensemble mean up to D+5. The red circles indicate that the difference in mean ACC is significant at the 5% level. From D+7, the ACC of the EMN is significantly larger than that of the DET forecast.
- For θ_2 (middle two curves) the cross-over occurs earlier, with the ACC of the EMN becoming significantly larger by D+5. Note that the T63 truncation of the Z_{500} and θ_2 data prior to this analysis is the primary reason for the larger ACC values here than in Figure 1.
- For $P_{p/c}$ (bottom set of curves) the ACC of the EMN is significantly larger than that of the DET forecast by D+3.

Hence it appears that as the inherent spatial scale of the field gets smaller, the cross-over occurs earlier. One can argue that this is a consequence of the EMN acting as an "intelligent filter" that removes features that are less predictable and thus more potentially harmful to the ACC. However, there are many problems associated with the EMN. For example the EMN will not, in general, even represent a dynamically valid atmospheric state beyond the first few days of the forecast (if non-linearity is important). Also it may well be that the user is particularly interested in the less predictable features (such as extreme events for example) that have been filtered out. Hence the deterministic comparison of the two forecast systems is less clear than it might seem at first sight (particularly after the cross-over has occurred).

Also plotted in Figure 2 (blue curve) is the ACC of European $P_{p/c}$ for the single T255 control forecast "CNT". This can be more legitimately compared with the DET ACC (red curve). It is clear that the lower-resolution forecast is significantly worse (indicated by the blue circles) than the higher-resolution forecast up to D+7. Beyond D+7, there is little to be gained with the higher resolution, at least in terms of this particular rainfall score.

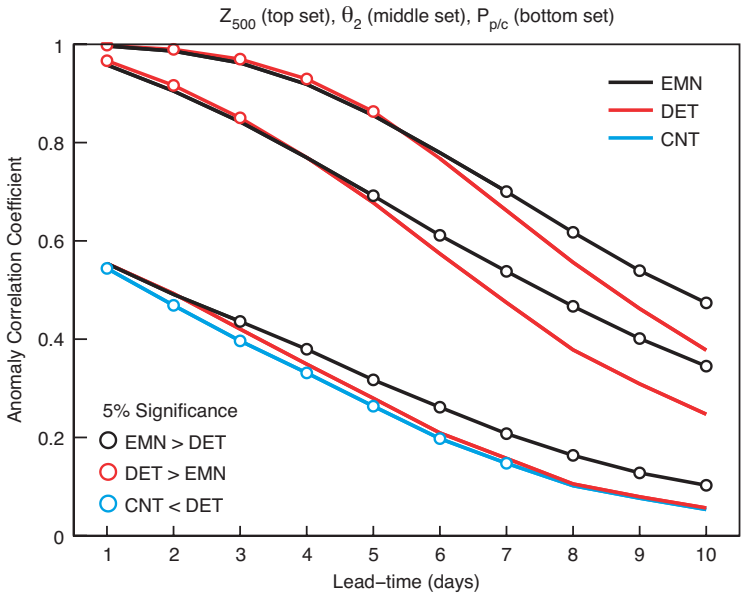


Figure 2 As Figure 1 but for Anomaly Correlation Coefficients (ACCs) of the DET forecast (red) and ensemble mean ("EMN", black) forecast. The Z_{500} (top two) and (θ_2 (middle two) curves are the mean ACCs of all 12 UTC forecasts made between 6 June 2004 and 5 June 2005. Z_{500} and θ_2 data are truncated to T63 and interpolated to a 1.875° regular grid before the ACCs are calculated. The $P_{p/c}$ (bottom) curves are the mean ACCs of all 12UTC forecasts over the four years 2001–2004. Also shown is the ACC of $P_{p/c}$ for the control forecast ("CNT", blue). Circles indicate a 5% statistically significant difference using a paired t-test taking autocorrelation into account. Red circles show where the DET forecast is significantly better than the EMN, black circles show where the ACC of the EMN is significantly larger than that of the DET forecast and blue circles show where the CNT forecast is significantly worse than the DET forecast.

doi:10.21957/cd347812th

Probabilistic comparison of deterministic and ensemble prediction systems

Probabilistic forecasts require the definition of an "event" (here we will choose the event that 24-hour accumulated precipitation exceeds 1 mm). A probabilistic score requires as input a probability that the particular event will occur. To get a probability from the EPS is simply a matter of counting the fraction of ensemble members that predict the event will occur (if all EPS members are equally likely). To get a probability from the DET forecast one possibility is to set the probability to 1 if the DET predicts the event will occur and set the probability to 0 otherwise (refinements will be discussed briefly later). The score that has been used here is the Brier Skill Score (BSS). The BSS is a measure of how well we forecast the probability that the event will occur relative to a forecast that simply uses the climatological probability. Clearly for a probability forecast at a single location and for a single date, it is not possible in general to determine the accuracy of the probability forecast. However, the Brier Skill Score is calculated here over many forecasts (4 years of daily forecasts) and over many station locations (typically 300–400 each day). Roughly speaking, the more often the event occurs when the forecast probability is high and the less often the event occurs when the forecast probability is low, the more positive will be the BSS. A perfect forecast system would have a BSS of 1. For further information about how the BSS used in this study compares with the present operational methodology see Box A.

Figure 3 (black curve) shows the EPS BSS for the prediction of the event that 24-hour accumulated precipitation exceeds 1 mm. The BSS is based on all 1461 of the 12 UTC forecasts made between 2001 and 2004. As before, the score refers to the skill in predicting the precipitation accumulated over the preceding 24 hours. For D+1 and D+2, the BSS is around 0.33. It is unclear at present why the BSS at D+2 is as good as at D+1. It is also unknown whether this feature occurs for other precipitation thresholds. Further investigation is planned. Beyond D+2, the BSS of the EPS declines but remains positive (and thus potentially useful) to D+9.

The BSS for the DET forecast is also shown in Figure 3 (red curve). At D+1, this is already lower than the BSS for the ensemble prediction system (black) and it drops very rapidly with increasing lead-time, becoming negative beyond D+3. One may argue, therefore, that in probabilistic terms, the DET forecast is much worse than the EPS. However, the DET forecast gives a dichotomous outcome (0 or 1 of the event occurring) and simple mathematics shows that the BSS for a perfect model DET forecast should tend to -1 as the lead-time increases. (Even with a perfect model chaos will still be present and thus predictability will be lost at some lead-time.) This is in contrast to the limiting value of 0 for a perfect-model large-ensemble probabilistic forecast. Forecast "dressing" (e.g. by using past verification data to determine a non-dichotomous probability of the event occurring as a function of the magnitude of DET forecast rainfall) could be used to reduce the decline of the BSS of the DET forecast (and also possibly the decline of the EPS). Hence again, comparison of the two systems is problematic and misleading.

Note that the BSS for the single CNT forecast is also shown in Figure 3 (blue curve). The blue circles signify that this is statistically significantly worse, at the 5% level, than the BSS for the deterministic forecast (red).

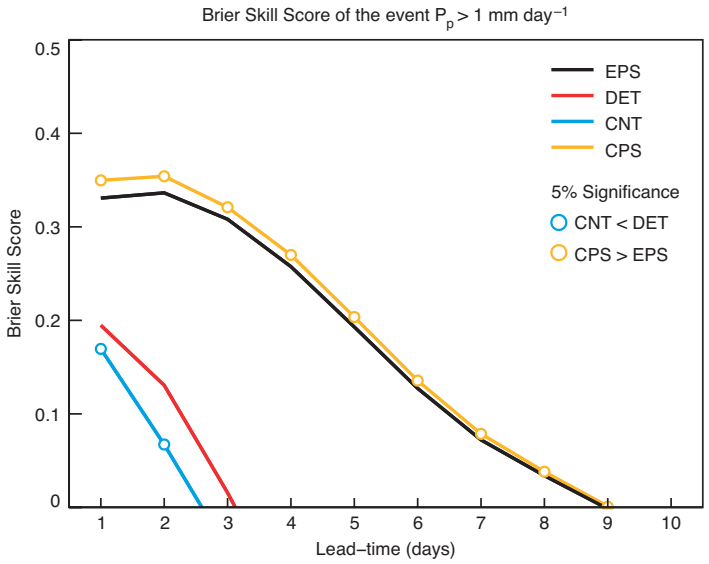


Figure 3 Brier Skill Scores (BSSs) for the event that daily-accumulated station-location point precipitation is greater than 1 mm (P_p>1): black for the ensemble prediction system ("EPS"), red for the DET forecast, blue for the CNT forecast, and orange for the combined prediction system ("CPS") which includes the EPS and the DET forecast. Blue circles indicate where the CNT forecast is significantly worse than the DET forecast and orange circles indicate where the CPS is significantly better than the EPS. The climatology used to calculate the BSS is based on a 1971–2000 observational database.

More about the Brier Skill Score **A**

The BSS for station-location precipitation is calculated operationally at ECMWF but there are differences with the method used in this study. Here, the climatological probability is a function of station location and derived from a long-term climate. The operational method assumes the climatological probability is not a function of location and is derived from the "sample climatology" (i.e. the rainfall that fell within the actual month being scored). In practice, particularly in summer, the climatological frequency of rainfall events varies greatly between the arid Mediterranean region and wetter northern Europe and thus, in this respect, the present methodology seems preferable. By using a sample climate, the operational method is likely to unduly penalise the forecast because the operational BSS is effectively scoring the model's ability to predict intra-monthly variability. Now that a long-term climatology is available for Europe, the present methodology seems preferable.

The operational score is also based on forecast data that is interpolated to a regular 1.5° grid before interpolation to station locations. The rationale for this is that it homogenises the record by reducing the sensitivity of the BSS to changing model resolution. Here, on the contrary, it is thought that the impact of increased resolution should be reflected in the skill score and interpolation is therefore done directly from the model grid to the station locations.

Finally, the operational scores are based on a fixed list of SYNOP stations. Again the rationale for this is to homogenise the record. Here it was thought that, for any particular day, all reporting stations should be used because it is possible that there will be missing data from any fixed list of stations and, over the years, stations may cease to exist and others may take their place. It is unclear which option is preferable.

Combined Prediction Systems: How to predict rainfall occurrence better

Given that it is problematic to directly compare the DET forecast with that of the EPS, the aim here is to take a very different approach – an attempt is made to combine the forecast information of both systems in order to get a better probabilistic forecast than either system can produce alone. The focus will be on precipitation because it is one of the most important quantities that we wish to forecast and, owing to its small-scale structure, there is a good possibility that the high-resolution forecast will be able to contribute substantially to the skill at short lead-times.

Figure 4 shows schematically how two forecast systems (ten-member ensemble and single deterministic) could be combined to give a single forecast probability for the event that the daily-accumulated precipitation will be greater than 1 mm. The figure shows a frequency plot of the forecast rainfall amounts from the individual ensemble members (orange squares) and the single high-resolution deterministic forecast (yellow rectangle). We have assumed that the high-resolution forecast should have the same weight as three ensemble members to reflect the possibility that it may be more skilful. The orange squares are all the same size because each ensemble member is equally likely. Based on the schematic, the combined forecast probability for the event that precipitation is greater than 1 mm is therefore 9/13. Combined (DET with 50-member EPS) probabilities for each day and each station are calculated in this way and used in the calculation of the BSS. In reality, we do not know beforehand the weight to apply to the DET forecast. We assume that the weight is a function of forecast lead-time but independent of station and, initially, independent of the time of year. Here, the weights are determined (from an analytical equation) so as to maximise the BSS. The weights determined for year n are used in the calculation of the BSS for a year n+1, thus avoiding any artificial enhancement of skill. Note that for the first year, 2001, the weights come from 2002 so are still independent of the forecast period.

Figure 3 (orange curve) shows the mean BSS for years 2001–2004 based on this "combined prediction system" (termed here, "CPS"). The orange circles signify where the CPS is statistically significantly superior at the 5% level to the ensemble system alone (based on daily contributions to the BSS). It is clear that the incorporation of the single high-resolution DET forecast improves the skill at all lead-times, particularly at short lead-times. This improvement in skill for all lead-times emphasises just how misleading is the dramatic drop of the BSS for the DET forecast alone. Further cross-validated tests reveal that the increased skill of the CPS occurs for every one of the 16 seasons in the study with perhaps the biggest increases occurring in autumn and winter.

doi:10.21957/cd347812th

Figure 5 shows the 2001–2004 mean of the optimal weights of the DET forecast as a function of lead-time. The weights determined for each year are actually very similar to those shown in Figure 5. At D+1, the DET forecast is equivalent to about 17 EPS members. Interestingly, this is roughly the number of T255 EPS members that could be made with the same computing power as a single T511 DET forecast (under operational configurations). At longer lead-times the weight of the high-resolution forecast diminishes so that by D+10, it is equivalent to about 2.5 ensemble members.

An important question to ask is whether the benefit of the high-resolution deterministic forecast comes from its high resolution or from the fact that it is initiated from a better estimate of the truth than any individual ensemble member. (The EPS members are initiated from our best estimate of the truth plus a small perturbation to reflect the uncertainty in our knowledge of the true atmospheric state.) To address the question of resolution versus initial conditions, the CNT forecast has been used instead of the DET forecast in the CPS. The CNT forecast is also started from our best estimate of the truth but run at the same resolution as each EPS member. When CNT is combined with EPS it is found that the optimal weight for the CNT forecast is very low (it is actually equivalent to -8 EPS members at D+1 and then tends to the weight of +1 EPS member by D+5). Hence it would seem clear that the benefit of the DET forecast comes predominantly from its higher resolution.

The negative initial weight for the CNT forecast requires further investigation. One hypothesis is that a negative weight is an efficient way of increasing the effective ensemble spread at short lead-times. (At short lead-times, the CNT forecast will lie close to the centre of the EPS distribution.) If this is the reason then the same effect should be occurring in the DET+EPS combination. The fact that the optimal weight for the DET forecast is +17 EPS members at D+1 may therefore imply that the high-resolution DET forecast is bringing rather more useful information to the CPS than may appear at first sight. This speculative explanation will be investigated in the future.

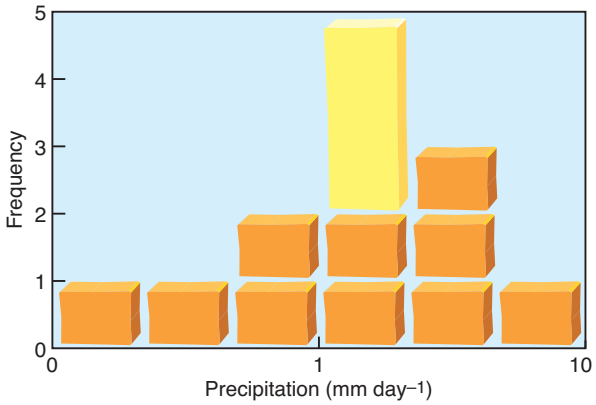


Figure 4 Schematic frequency diagram showing how the deterministic and ensemble forecast systems are combined to produce a single forecast probability for the event that 24-hour accumulated precipitation is greater than 1 mm. The orange squares represent individual ensemble members and the yellow rectangle represents the single high-resolution deterministic forecast. See the main text for further details.

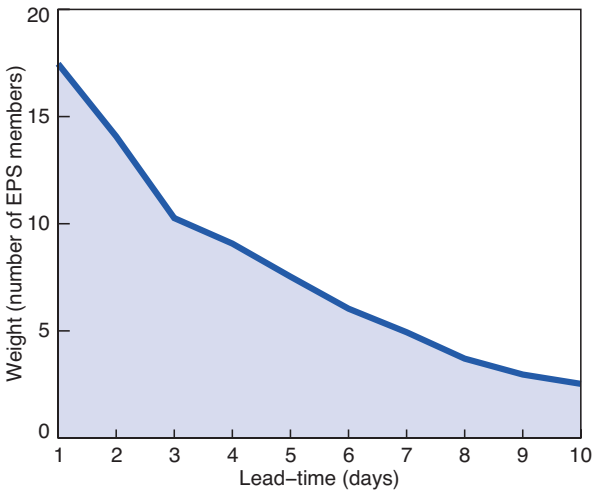


Figure 5 The optimal weight to apply to the high-resolution deterministic forecast (DET) in order to maximise the BSS in the combined prediction system (CPS) for the rainfall event that P_p>1 mm. The unit is the weight of one EPS member. Weights are applied in a cross-validated manner.

Further experiments with the Combined Prediction System and extensions

Further tests were made to see if adding a seasonal dependence to the weights applied to the CPS could increase overall skill. No improvement (or degradation) was found. One possibility is that there is little seasonal dependence. Another possibility is that the reduction in available training data leads to poorer estimates of the weights and this balances improvements arising from seasonal dependence.

There is a lot of interannual variability in the BSS (for the event P_p>1). For example 2003 has a value of around 0.4 for D+1 and D+2 while 2004 has values of around 0.3. The interannual variability does not appear to be associated with changes in ensemble design (e.g. the tuning of the spread) as the same variability is apparent in the BSS for the DET forecast. Since the ACC for precipitation is actually higher at D+1 and D+2 for 2004 than for 2003, it also seems unlikely that the model deteriorated in this respect between these two years. Hence it is possible that the BSS for the event P_p>1 is highly flow-dependent. The high 2003 score is predominantly due to high scores in spring and summer and could be associated with the European heat wave/drought that occurred at that time.

The method of combining forecasts has been extended to give the option of incorporating a third forecast system. The natural choice here would be to combine the EPS, DET and CNT forecasts. However, as with just the EPS and CNT combination, the optimal weights for the CNT are negative. Nevertheless, the option of combining three (probabilistic or deterministic) forecast systems may be useful in the future.

Future prospects

We have seen that combining the high-resolution deterministic forecast and ensemble prediction system in an optimal way can lead to better probabilistic rainfall forecasts for Europe than either system alone. The "Combined Prediction System" (CPS) benefits from the high-resolution attribute of the deterministic forecast (DET) at short lead-times and the probabilistic attribute of the ensemble prediction system (EPS) at longer lead-times. Both of these attributes are incorporated in the variable resolution EPS ("VAREPS") system which is shortly to be implemented at ECMWF. VAREPS, which has a higher resolution early in the forecast and is truncated to a lower resolution later on, presents a good framework for the prediction of rainfall. Results such as those presented here could help determine the best configuration of resolutions and truncation time for VAREPS.

Clearly other skill scores and other definitions of the weather "event" to be forecast could be analysed in a similar way. Indeed different precipitation thresholds are being investigated at present. These alternatives may produce different optimal weights for the systems combined within the CPS. In the article by *Palmer et al.* in this edition of the Newsletter there is a discussion on temperature prediction using the EPS and DET forecast. They apply a "dressing" to their forecasts which, as noted above, can improve probabilistic skill scores. Although this approach may clearly be useful in practice, here no dressing is done and the optimal weights are defined to be independent of location. The reason for this is that our goal is not to optimise predictability for specific (SYNOP) locations but to assess the typical skill for any locations (even where calibration data is not available). Hence BSS for the rainfall event P_p>1 mm is, in theory, applicable at any point and not just at SYNOP station locations.

A highly used product of ECMWF is the "meteogram". These meteograms display, for any desired location, DET and lower-resolution EPS control (CNT) forecasts of cloud-cover, precipitation, wind speed and temperature together with the quartiles of the EPS distribution. A difficulty for some users is to decide whether to "believe" the deterministic or probabilistic forecast. One could imagine giving the user the choice of an alternative meteogram that simply displays a CPS probability distribution. Tests would be required to see if the optimal weights are sensitive to the choice of threshold (1 mm, 5 mm of rainfall, etc.) and if a simple dressing of the DET forecast would be beneficial.

Further reading

Palmer, T.N., R. Buizza, R. Hagedorn, A. Lawrence, M. Leutbecher & L. Smith: Ensemble prediction: A pedagogical perspective. In this issue.

Simmons, A.J. & A. Hollingsworth, 2002: Some aspects of the improvement in skill of numerical weather prediction. *Q.J.R. Meteorol. Soc.,* **128,** 647–677.

 doi:10.21957/cd347812th