ECMWF Feature article

from Newsletter Number 107 – Spring 2006

METEOROLOGY

Limited area ensemble forecasting in Norway using targeted EPS

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

doi:10.21957/op5410xa7y

This article appeared in the Meteorology section of ECMWF Newsletter No. 107 – Spring 2006, pp. 23–29.

Limited area ensemble forecasting in Norway using targeted EPS

Marit Helene Jensen, Inger-Lise Frogner, Trond Iversen, Ole Vignes
(Norwegian Meteorological Institute)

The forecasting of severe weather is a high priority for national weather services and weather prediction centres. Extreme weather events outside the tropics often occur over small areas and last a short time. They are frequently micro- to meso-scale structures embedded into larger meso-scale and synoptic-scale phenomena. Some small-scaled structures can be associated with geographically fixed features like topography, coastlines, sea-ice, and thermal contrasts, whilst others are caused by internal non-linear dynamics which produce sharp fronts, squall lines, etc. Upscale cascading of atmospheric un-predictability, from the range of hours for convective systems to a week or two for planetary scales, renders the prediction of pure dynamically-produced structures a considerable challenge. Structures associated with geographically fixed forcing, however, leave more hope for the forecaster (*Anthes et al.,* 1985; *Boer,* 1994). There are reasons for optimism if a high quality prediction of large meso- and synoptic-scale phenomena can be combined with a better representation of the lower boundary forcing.

Hazardous weather is, by nature, unusual. A probabilistic approach is particularly appropriate for predicting such weather. The existing global ensemble prediction systems (e.g. the ECMWF Ensemble Prediction System – EPS) presently have considerable probabilistic skill on synoptic and large meso-scales. However, a number of meso-scale processes are not well captured, and the fixed boundary forcing will frequently be too smooth with the present resolution. A well-designed high-resolution, limited-area ensemble prediction system (LAMEPS) using a skilful synoptic-scale EPS at the open boundaries should improve this situation. Running a LAMEPS provides an ensemble with higher horizontal resolution. Furthermore, limiting our forecast interest to a smaller target domain enables a large fraction of the prediction spread to be covered with a small ensemble size.

At the Norwegian Meteorological Institute we have run such a LAMEPS quasi-operationally since mid-February 2005. LAMEPS is run with the Norwegian version of the HIRLAM model and it is driven by members of the ECMWF EPS which are targeted to produce maximum spread amongst ensemble members after 48 hours in northern Europe and adjacent sea areas. This system is abbreviated to targeted EPS, or simply TEPS. We have made a comparison between the 50 member EPS and the 20 member TEPS for an area covering much of the target area. The TEPS system is described in *Frogner & Iversen* (2001).

Pre-operational studies of LAMEPS in Norway have shown promising results. For further information see the article by *Frogner et al.* (2005).

A multi-model ensemble system (NORLAMEPS) is also used which simply combines LAMEPS and TEPS by using all ensemble members from both systems simultaneously. This combination gives a larger ensemble without extra model runs. Even though the combined system is to some extent an auto-duplication, the ensemble spread is larger for two reasons: there are un-correlated differences between fields from the different models, and the LAMEPS control forecast with HIRLAM can deviate considerably from the TEPS control with the ECMWF Integrated Forecast System (IFS).

Here we describe the model setup for LAMEPS, TEPS and NORLAMEPS, and show some verification results for the summer and spring of 2005. We also discuss different verification methods before we give our conclusions and suggest future work.

Brief descriptions of the various systems used in the investigation are given in Table 1.

doi:10.21957/op5410xa7y

Abbreviation	Description
EPS	ECMWF Ensemble Prediction System using 50+1 ensemble members
LAMEPS	Limited-area ensemble prediction system using 20+1 ensemble members based on the Norwegian version of the HIRLAM model
TEPS	Targeted version of EPS using 20+1 ensemble members with Northern Europe and adjacent sea areas as the target area
NORLAMEPS	Multi-model ensemble system using 41+1 ensemble members which combines LAMEPS and TEPS

Table 1 Descriptions of the various systems used in the investigation of limited area ensemble forecasting.

Model setup

LAMEPS is an ensemble of runs with the Norwegian version of the limited area model HIRLAM (horizontal resolution of 0.2° with 40 levels). It uses ensemble members from TEPS to perturb both the initial and the lateral boundary conditions.

TEPS uses the same model version and the same set-up as used for the operational EPS. Only 10 singular vectors are calculated as opposed to 25 in the EPS. These singular vectors are targeted to maximize the total energy at final optimization time (48 h) in Northern Europe and adjacent sea areas (see Figure 1). Singular vectors at initial time and 48 hours evolved singular vectors valid at the same time are combined to form initial state perturbations. These are added to and subtracted from the initial state analysis (the "control") with amplitudes based on estimates of analysis error. TEPS thus contains 20 ensemble members in addition to the control forecast. The TEPS forecast length is 72 hours and is run at ECMWF once per day at 12 UTC.

Each LAMEPS ensemble member is constructed by running HIRLAM from 20 alternative initial states obtained by adding the 20 time-developed (6 hours forecasts) TEPS ensemble perturbations (the difference between each TEPS ensemble member and the TEPS control) to the HIRLAM 18 UTC analysis. At the open lateral boundaries the time-developed TEPS ensemble members, corresponding to those used for initial perturbations, are imposed. Thus we obtain 20 different forecasts in addition to the HIRLAM control run. Since HIRLAM starts with an 18 UTC analysis and TEPS with a 12 UTC analysis, the forecasts from LAMEPS are 6 hours shorter than the forecasts from TEPS and EPS. LAMEPS is run at 18 UTC every day and the forecast length is 60 hours.

NORLAMEPS combines the forecasts from TEPS and LAMEPS to provide a single statistic for events, even though they are not entirely independent of each other. Without extra cost, the total number of ensemble members is then 41 in addition to the HIRLAM control forecast. In this way NORLAMEPS is supposed to partly account for uncorrelated forecasts errors caused by model imperfections. The differences between the initial fields in TEPS and LAMEPS are partly caused by these model differences.

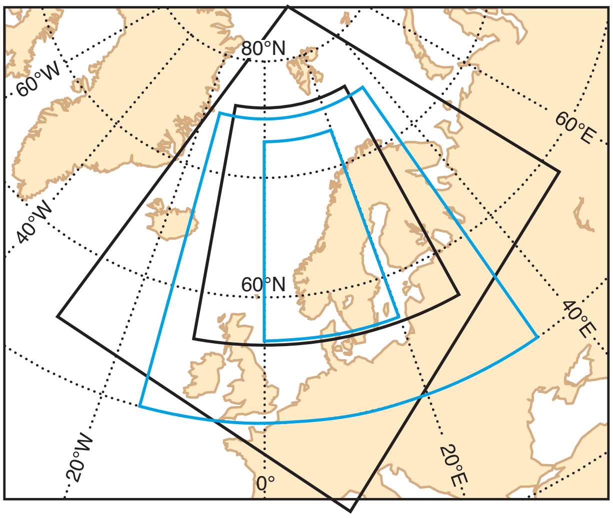

Figure 1 Areas used in the experiments. The large black area is the HIRLAM integration area, the large blue area is the target area for optimization of singular vectors, and the small blue area is the small verification domain and the small black area is the verification domain for the EPS/TEPS comparison.

Verification methodology

Verification of precipitation forecasts against SYNOP observations is not straightforward because of the very different scales of observations and the forecasts. Furthermore, LAMEPS and TEPS/EPS also have different resolutions. Hence comparison of the different systems using either the analysed fields from the HIRLAM or the ECMWF IFS model as “the truth” would favour one of the systems. To overcome this problem we use the approach proposed by *Ghelli & Lalaurette* (2000) and construct so-called “super-observations” which are representative of precipitation grid-squares. Here all precipitation stations in Norway (several hundreds) inside the verification area (Figure 1) are aggregated to regular grids. Precipitation “super-observations” representative of our 0.2°x0.2° HIRLAM grid are thus calculated. All the verification of precipitation described here uses such super-observations.

Total precipitation from LAMEPS, TEPS, EPS and NORLAMEPS are compared to the super-observations using Rank Histograms, Reliability Diagrams, Brier Skill Scores, ROC curves, and cost/loss analysis (*Katz & Murphy,* 1997). The diurnally accumulated precipitation observations are taken at 06 UTC. Since the forecasts starts at 12 UTC and 18 UTC and are 66/60 hours long, this leaves only two possible time-intervals in the forecast range for verification: (+12/18 h to +36/42 h) and (+36/42 h to +60/66 h). Note that since LAMEPS is started 6 hours later than TEPS and EPS, the forecast from LAMEPS is 6 hours shorter than the other two forecasts.

The distribution of precipitation in Norway is dominated by sharp gradients, caused by predominant westerly winds, a long coastline, and a complex topography. The gradient across the divide between the western and eastern watersheds in Southern Norway is particularly large: western watersheds receive large amounts of precipitation, while the eastern ones are frequently sheltered by the mountains. Typically the annual difference amounts to a factor of 2 to 3 (Figure 2), but in several cases the differences are even much larger. It was noted by *Hamill* (2005) that agglomerations of samples spanning locations and times with different climatological frequencies can lead to spurious skill measures. To circumvent this problem we verify separately three sub-regions with grossly different precipitation climatologies (Figure 2). The precipitation frequencies also vary over the year, and we split the verification results into spring (February-April) and summer (May-July). Averages are calculated using weights reflecting the area of the sub-regions and by the number of days in the two periods.

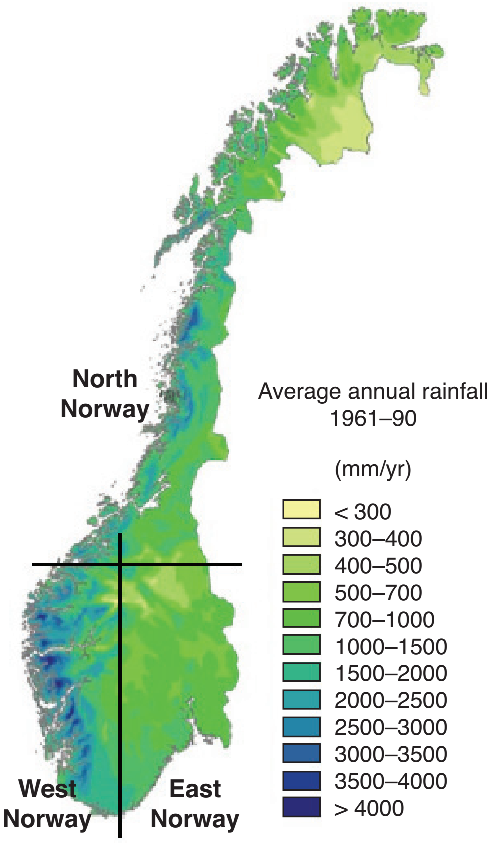

Figure 2 Average annual precipitation amounts in Norway from 1961-90 (mm). Also shown are the three different verification areas based on precipitation frequency.

The ensemble spread

The spread of an ensemble forecast can be used as an indicator of quality – see Box A for the definition of the ensemble spread.

Table 2 shows the spread of total precipitation forecasts for each ensemble system. The spread increases with forecast time, which is in line with expected behaviour of unstable systems starting from small perturbations. The EPS has the smallest spread for both forecast times. The main reason for targeting is to constrain the perturbations to a predefined area of particular interest over a certain forecast range. It is expected that a targeted system will have a larger spread between ensemble members in this target area than a system that is not targeted, given the same number of singular vector based perturbations. Thus a 20-member TEPS ensemble has a considerably larger spread between the members than EPS with 50 members. Hence the TEPS ensemble includes a wider selection of fast-growing disturbances over the time-range and area of interest than the EPS ensemble. As a consequence forecasts made from TEPS over the selected time-range can quantify risks of more extreme cases better than those based on EPS. The realism of the increased risk needs to be investigated.

The EPS ensemble with 50 members has considerably smaller spread than both ensembles from TEPS and LAMEPS, each of which has only 20 members. The LAMEPS ensemble has a slightly larger spread than the original TEPS ensemble, even though the initial and boundary perturbations are entirely based on TEPS.

NORLAMEPS, which is a simple combination of the TEPS and LAMEPS, has the largest ensemble spread of all the systems for forecast range 36/42 h. Hence, LAMEPS triggers slightly different unstable structures in HIRLAM than the global model used for generating TEPS. This difference can partly originate from the higher resolution of LAMEPS and partly from the fact that two different models are used to compute the ensemble members in NORLAMEPS. The spread in the NORLAMEPS ensemble and the associated risks of potentially extreme weather developments should therefore be taken as due to a combination of chaos and model uncertainty. It is impossible to tell to what extent the additional spread from model uncertainty stems from uncorrelated model errors or if it is due to equally realistic but different time developments.

	LAMEPS	TEPS	EPS	NORLAMEPS
+36/+42	2.15	2.08	1.56	2.19
+60/+66	2.47	2.38	2.07	2.47

Table 2 Spread around ensemble mean (in mm/day) for total precipitation for the four ensemble systems.

Definition of ensemble spread

A

The spread is the rms-difference between the ensemble members and the ensemble mean defined as:

$$S = \frac{1}{I} \sum_{i=1}^{1} \sqrt{\frac{1}{N \cdot D} \sum_{n=1}^{N} \sum_{d=1}^{D} \left(e_{ind} - m_{id} \right)^2}$$

where I is the number of grid-points inside the verification area, N is the number of ensemble members, D is the number of cases, e_{ind} is the ensemble member value for member n in the case d and at the specific point i, and m_{id} is the ensemble mean for the same case and at the same point.

A small spread is an indication of high skill. However, a large spread does not necessarily indicate high skill though it does give an indication of low predictability.

Skill of the four ensemble systems

All four ensemble systems are evaluated for total precipitation accumulated over the two verification periods using several measures of skill. The observations of precipitation are mainly at 06 UTC and therefore we use this time for the verification. The verification period for LAMEPS differ from that for TEPS and EPS because of the forecast starts 6 hour later for LAMEPS. The verification periods are therefore 12–36 h and 36–60 h for LAMEPS, 18–42 h and 42–66 h for TEPS/EPS and a combination of these periods for NORLAMEPS. The scores are calculated separately for each of the three sub-domains shown in Figure 2 and for the two seasons (spring and summer). The area-weighted average over the domains and the two seasons are computed.

Figure 3 shows the Brier Skill Scores for the 12/18–36/42 h and 36/42–60/66 h forecasts as a function of precipitation threshold. The Brier Skill Score measures the improvement of the probabilistic forecast relative to climatology with 0 indicating no skill compared to climatology. Figure 3 shows that LAMEPS has considerably lower scores than the other systems for both forecast periods, especially for small precipitation amounts. For 12/18–36/42 h TEPS has higher scores than EPS up to around 20 mm/day, but for larger thresholds EPS has higher scores. During 36/42–60/66 h TEPS has higher scores than EPS for all thresholds. NORLAMEPS has higher scores than any of the other systems for the high precipitation amounts (10–15 mm/day) for both forecast periods.

The ROC curve is a plot of the hit rate against the false alarm rate and it gives an indication of the ability to distinguish between events and non-events. A measure of skill is the area under the curve which has a maximum value of 1 with no skill being indicated by a value of 0.5. Figure 4 shows the area under the ROC curves for the various precipitation thresholds for the four ensemble systems. TEPS is comparable with EPS for both verification periods with EPS a bit better. For low thresholds LAMEPS has clearly lower scores than for the other systems. NORLAMEPS has the highest scores for mid to high thresholds.

For both Brier Skill Scores and area under ROC curves the 20 members of TEPS and the 41 members of NORLAMEPS give as good or better results than EPS with 50 members.

An alternative way of assessing the quality of probability forecasts is to consider the relative improvement in economic value (Relative Value) as a function of the cost/loss ratio with climatology as the reference. As with the other skill scores, a Relative Value of 1 indicates a set of perfect forecasts. Figure 5 shows the Relative Value for the weighted mean over the three areas and the two seasons with an event threshold of 5 mm/day. For the Relative Value the EPS gets the highest values, but NORLAMEPS has the widest distribution.

Note that all three measures of skill indicate that LAMEPS has a considerably lower score for the low precipitation thresholds. For mid to high thresholds NORLAMEPS has very good scores, showing that LAMEPS gives extra and valuable information to the TEPS.

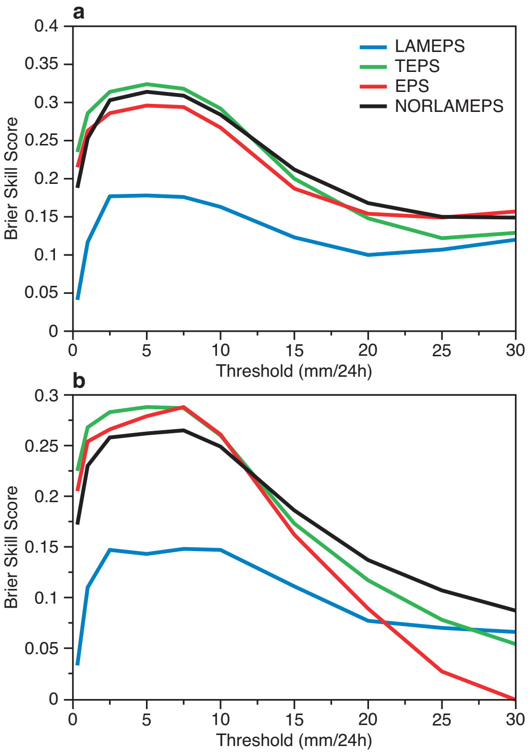

Figure 3 Brier Skill Score for precipitation as a function of threshold for LAMEPS, TEPS, EPS and NORLAMEPS for (a) 12/18-36/42 h and (b) 36/42-60/66 h forecasts. Mean over all verification areas and both seasons.

We have also looked at the Rank Histograms and the Reliability Diagrams for the four systems (not shown here). The Rank Histograms indicate a bias in all four systems where they all underestimate the variability. The underestimation is small in TEPS and especially large in EPS. The Reliability Diagrams show good reliability up to about 70% for TEPS and NORLAMEPS, after which the two systems over-forecast the higher probabilities. LAMEPS and EPS over-forecast the probabilities from about 30–40%.

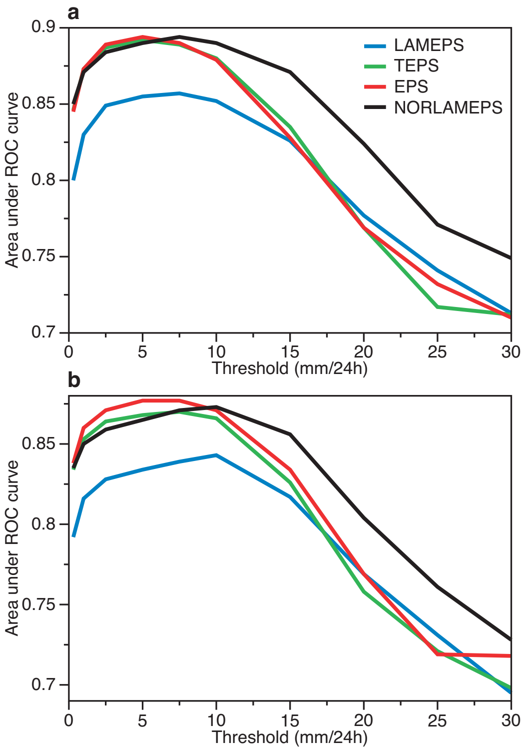

Figure 4 Area under the ROC curve for precipitation as a function of threshold for LAMEPS, TEPS, EPS and NORLAMEPS for (a) 12/18–36/42 h and (b) 36/42–60/66 h forecasts. Mean over all verification areas and both seasons.

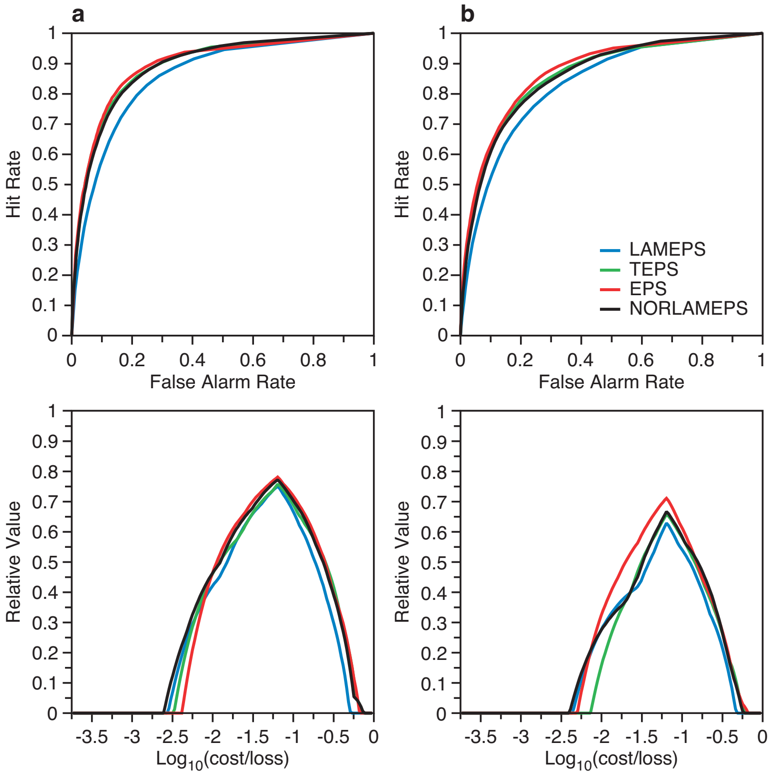

Figure 5 (a) ROC curves (top) and Relative Value analysis (bottom) for total 24 hour precipitation for LAMEPS, TEPS, EPS and NORLAMEPS for 12/18–36/42 h forecasts. (b) As (a) but for 36/42–60/66 h forecasts. The threshold is 5 mm/day. Mean over all verification areas and both seasons.

Additional comparison between EPS and TEPS

The precipitation observations used to compare the four ensemble systems cover only a small part of the targeting domain. Therefore an additional investigation was carried out to study the differences between the EPS and TEPS for a larger targeting domain. The aim was to assess whether there are benefits in the targeting for larger areas than Norway. This should be interesting for others wanting to use TEPS, both in itself and as an input to limited area models. Given the different ensemble size of EPS and TEPS, we have compared the 20 member TEPS against both the 50 member EPS and against only the 20 first members of the EPS.

For this larger area we did not have observations of precipitation, so we have used 30 hour forecasts of 24 hours accumulated total precipitation from the deterministic ECMWF model (T511L60) as the truth. The verification area covers much of Scandinavia and the sea areas west of Norway (the small black area shown in Figure1).

We have used the same 0.2 x 0.2 degree grid as used earlier in the study, and the statistical scores are the same as for the verification against observations.

Here we display the same probabilistic scores that we used for the previous comparisons: Brier Skill Scores (Figure 6) and area under the ROC curves (Figure 7) for different precipitation thresholds, and ROC curves and cost/loss analysis (Figure 8) for an event threshold of 5 mm/day.

The results given in Figures 6, 7 and 8 can be summarised as follows.

- For all measures of skill, the full 50 member EPS has, as one should expect due to the higher ensemble members, higher scores than the 20 member EPS as well as the 20 member TEPS.
- For the Brier Skill Score, the 20 member EPS has a somewhat higher score than TEPS for small and medium precipitation amounts, but for high precipitation amounts TEPS has a higher score than the EPS with the same number of members.
- For the Brier Skill Score, the scores for TEPS is closer to the scores from EPS for the second verification period, but for the area under the ROC curve the scores for TEPS are better compared to the others for the first verification period.
- For the other verification parameters it seems as if TEPS is as good as or better than the low member EPS for the first verification period.

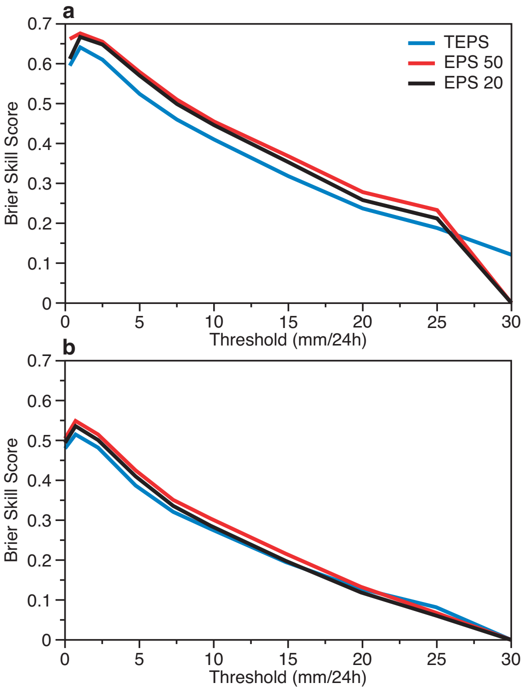

Figure 6 Brier Skill Score for precipitation as a function of threshold for TEPS, 50 member EPS (EPS 50) and 20 member EPS (EPS 20) for (a) 18–42 h and (b) 42–66 h forecasts.

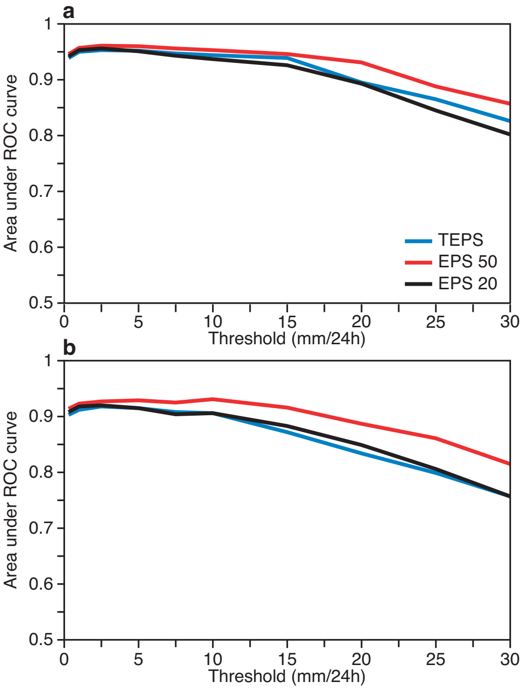

Figure 7 Area under the ROC curve for precipitation as a function of threshold for TEPS, 50 member EPS (EPS 50) and 20 member EPS (EPS 20) for (a) 18–42 h and (b) 42–66 h forecasts.

doi:10.21957/op5410xa7y

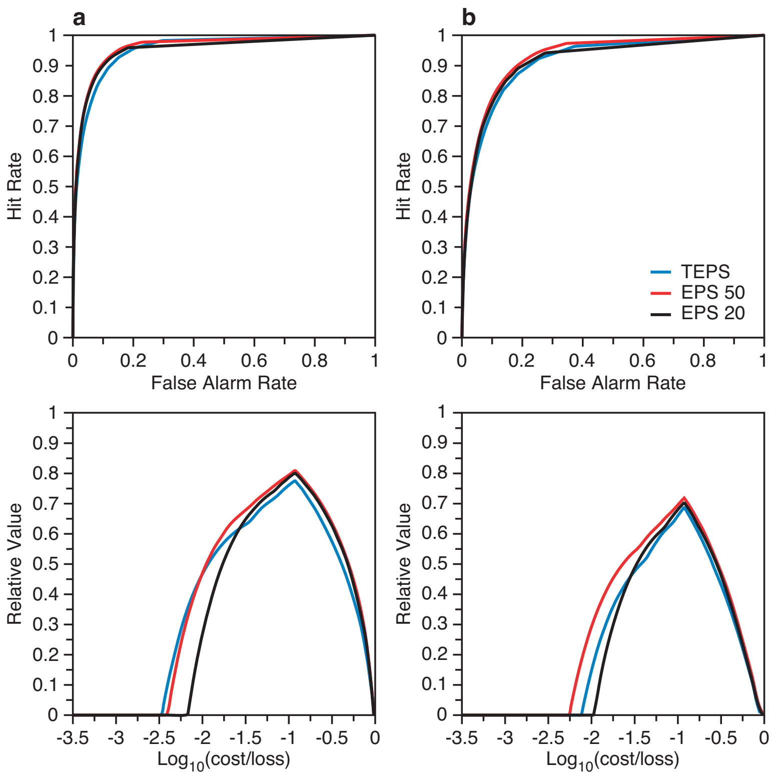

Figure 8 (a) ROC curves (top) and Relative Value analysis (bottom) for total 24 hour precipitation for TEPS, 50 member EPS and 20 member EPS for 18–42 h forecasts. (b) As (a) but for 42–66 h forecasts. The threshold is 5 mm/day.

Discussion and conclusion

From the results we have seen that LAMEPS is able to produce more spread than EPS for precipitation over Norway. For events with small precipitation amounts the probabilistic scores for TEPS and EPS are better than for LAMEPS, but for larger precipitation amounts LAMEPS scores better.

The combined system NORLAMEPS gets the largest spread and also best probabilistic scores from mid to high precipitation amounts. This shows that combining LAMEPS and TEPS adds value to the two individual systems. The improvement by NORLAMEPS can partly be due to the increase in resolution for LAMEPS, and partly from the fact that it combines results from two model systems with different characteristics.

For the Norwegian verification area the comparison of the 20 members TEPS and 50 members EPS provides some interesting results. The increase in ensemble spread of TEPS compared to EPS demonstrates the advantage of targeting. Also for many of the probabilistic scores TEPS is better or comparable to EPS, even though it has fewer members.

Using the larger verification area for the comparison one clearly sees the advantages of having an ensemble with more members. However, one can also see that the targeting gives extra value to the ensemble forecast of precipitation, especially for the shorter-range forecast.

The verification has shown that TEPS gives better results for the short-range verification period. This may indicate that the method of perturbing TEPS gives very good results early in the forecast range, but that the weather system then moves out of our target area. In this case the ordinary EPS has an advantage. One way of dispensing with this is to combine TEPS and EPS as input to LAMEPS.

Ongoing work

Our plans are to continue to develop LAMEPS/NORLAMEPS by, for example, perturbing the physics in HIRLAM model and increasing the resolution of LAMEPS (first to 15 km and then to 10 km). Experience with the operational HIRLAM models shows that HIRLAM with 10 km resolution is capable of predicting polar lows not resolved at 20 km resolution. As polar lows are important for our area of interest and often give intense precipitation and strong winds, it is desirable that LAMEPS should be able to resolve these features. The focus will also be extended to include more weather parameters such as temperature and wind.

Different ways of making perturbation for TEPS will be tested together with staff at ECMWF. Indeed we are grateful that the computing resources necessary to run TEPS have been partly provided by an ECMWF Special Project.

Further reading

Anthes, R.A., Y.H. Kuo, D.P. Baumhefner, R.M. Errico & **T.W. Bettge,** 1985: Predictability of mesoscale motions. Issues in atmospheric and oceanic modelling. In *Advances in Geophysics: Part B, Volume 28.* Eds. B. Saltzman and S. Manabe. Academic Press, Orlando, USA.

Boer, G.J., 1994: Predictability regimes in atmospheric flow. *Mon. Wea. Rev.,* **122,** 2285–2295.

Frogner, I-L. & **T. Iversen,** 2001: Targeted ensemble prediction for northern Europe and parts of the North Atlantic Ocean. *Tellus,* **53A,** 35–55.

Frogner, I-L., H. Haakenstad & **T. Iversen,** 2005: Limited area ensemble predictions at the Norwegian Meteorological Institute. Accepted for publication in *Q.J.R. Meteorol. Soc.*

Ghelli, A. & **F. Lalaurette,** 2000: Verifying precipitation forecasts using upscaled observations. *ECMWF Newsletter No 87,* 9–17. Available from ECMWF, Shinfield Park, Reading, RG2 9AX, UK.

Hamill, T.M., 2005: Common forecast verification metrices can overestimate skill. Submitted to *Mon. Wea. Rev.*

Katz, R.W. & **A.H. Murphy,** 1997: *Economic Value of Weather and Climate Forecasts.* Cambridge University Press, 222 pp.

Ripley, B.D., 1981: *Spatial Statistics.* Wiley, USA, 252 pp.

European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, RG2 9AX, England

doi:10.21957/op5410xa7y